Aus dem Max-Planck-Institut für Molekulare Pflanzenphysiologie
Abteilung Molekulare Pflanzenphysiologie
(Direktor: Prof. Dr. rer. nat. Lothar Willmitzer)

in Potsdam-Golm

Applied Metabolome Analysis: Exploration, Development and Application of Gas Chromatography-Mass Spectrometry based Metabolite Profiling Technologies

Habilitationsschrift
zur Erlangung des akademischen Grades Doctor rerum naturalium habilitatus
(Dr. rer. nat. habil.)
im Wissenschaftsfach
„Molekulare Physiologie“

eingereicht an der Mathematisch-Naturwissenschaftlichen Fakultät der
Universität Potsdam

von
Dr. rer. nat. Joachim Kopka,
geboren am 4. Mai 1962 in Münster (Westfalen)

Potsdam-Golm, 2008

Published online at the
Institutional Repository of the University of Potsdam:
URL http://opus.kobv.de/ubp/volltexte/2010/4059/
URN urn:nbn:de:kobv:517-opus-40597
http://nbn-resolving.org/urn:nbn:de:kobv:517-opus-40597

TABLE OF CONTENTS

INTRODUCTION

The metabolome - in analogy to higher system levels, namely, the genome, transcriptome and proteome - is defined to represent the complete metabolic complement of biological systems. The metabolome comprises an immense diversity of chemical compounds, ranging from gases, for example O_2 or CO_2, to polar or lipophilic small molecules and to polymers, such as starch. In contrast to the clearly delimited genome, the size of the metabolic complement present in a biological system can only be estimated. Genome based reconstructions of metabolic pathways and comprehensive screening of the current biochemical knowledge-base predict ~600 metabolites to be present in the unicellular yeast, *Saccharomyces cerevisiae* (Forster et al. 2003), or ~750 metabolites in *Escherichia coli* (Nobeli et al. 2003), and list about 1,500 metabolites for the human metabolome (Duarte et al. 2007). In the plant kingdom a biosynthetic potential of ~200,000 metabolites of highly diversified secondary metabolic pathways may be expected (Hall et al. 2002, Fiehn 2002, Fernie et al. 2004). In conclusion, the metabolome is best represented and understood as a complex network of metabolite pools which are linked by enzymatic or non-enzymatic reactions and communicate across system borders through facilitated transport and diffusion processes.

Late in the 1990s the metabolomics concept emerged independently in the fields of yeast (Oliver at al. 1998, Raamsdonk et al. 2001, Stephanopoulos et al. 2004, Nielsen and Oliver 2005), *Escherichia coli* (Tweeddale et al. 1998) and plant molecular physiology (Trethewey et al. 1999, **Fiehn et al. 2000a [1]**, **Roessner et al. 2000 [2]**, Roessner et al. 2001a, Hall et al. 2002)[*]. Shortly afterwards, the same concept was given the synonymous name, metabonomics, for human, clinical or toxicological applications (Nicholson et al. 1999, Nicholson et al. 2001, Lindon et al. 2005). Both concepts, metabolomics and metabonomics, are defined as the application of comprehensive metabolite analysis in the sense of a large scale phenotypic screening to aspects of functional genomics and molecular physiology. In the following the term metabolomics will be used.

Thus, at the turn of the century, the fourth conceptual and integral part of the Rosetta stone for modern systems biology was firmly put in place. However, it became immediately apparent that the current analytical tools for the monitoring of the metabolic complement were far from comprehensive. Due to the high chemical diversity of metabolites (e.g. Sumner et al.

[*] Publications relevant for this thesis are indicated by bold type and are numbered in square brackets according to their appearance within the appendix of publication facsimiles. Authored or co-authored publications not added to the appendix are underlined.

2003) and the large dynamic range of concentrations at which metabolites may occur in biological systems, exceeding 5,000-fold changes in extreme cases (e.g. van den Berg et al. 2006), traditional quantitative analysis was predominantly targeted at single or few chemically similar metabolites. In contrast, the metabolomics field turned to pre-existing multi-parallel analytical tools and to the investigation of complex metabolite preparations for the implementation of the underlying visionary concept of comprehensive analysis. Instead of exact quantification of metabolite pools, requiring metabolite specific quantitative calibration of analytical instruments, the estimation of relative changes of pool sizes was accepted for the large scale screening of samples (**Fiehn et al. 2000a [1]**, **Roessner et al. 2000 [2]**). Two variants of these screenings were added to the metabolomic tool box, namely, metabolite fingerprinting and metabolite profiling (e.g. Fiehn 2002). Fingerprinting is defined as the non-targeted analysis of all recorded signals of a given analytical technology without knowledge about metabolite identity, whereas metabolite profiling is typically restricted to the subset of analytical signals which can be linked to a broad and known set of pre-defined chemical compounds. Thus, a metabolite profile can be interpreted as a metabolic phenotype (Roessner et al. 2001a) and may support rational genetic engineering (Trethewey 2004). This seemingly simple leap of concept in metabolic analysis and the versatility of adapting respective analytical technologies to a vast range of biological systems led to the fast establishment of the metabolomics field. In retrospect, the metabolomics field has taken an astonishing and still highly dynamic development, as indicated by publication statistics (e.g. Guy et al. 2008a) and the rapid succession of foundations, such as the company Metanomics GmbH in October 1998, the Plant Metabolomics Society in 2002, the Metabolomics journal in 2005, and the Metabolomics Society including all biosciences in the same year (Fig. 1).

Indeed numerous analytical technologies have been applied in the metabolomic field. All have in common the potential of multi-dimensionality and the hyphenation of high-resolution separation to multi-channel detection. Two dimensional thin layer chromatography (2D-TLC) or paper chromatography, which led to the Nobel prize winning findings of Calvin (1962) and to the discovery of the photosynthetic dark reactions, may be seen as the first and, perhaps, so far most important application of early metabolic profiling. 2D-TLC has been revisited by modern metabolomics (Tweeddale et al. 1998). However, this technology was superseded by the application of gas chromatography-mass spectrometry (GC-MS), liquid chromatography-mass spectrometry (LC-MS), capillary electrophoresis-mass spectrometry (CE-MS) and nuclear magnetic resonance spectroscopy (NMR).

Each of the above and other tested technologies allows insight only into a limited metabolic window which is restricted by the respective analytical technology. As a consequence, the metabolome is currently best monitored by a combination of profiling methods, which may either be assembled in a company environment or by an academic consortium of experts in specific profiling techniques (e.g. Böttcher et al. 2008).

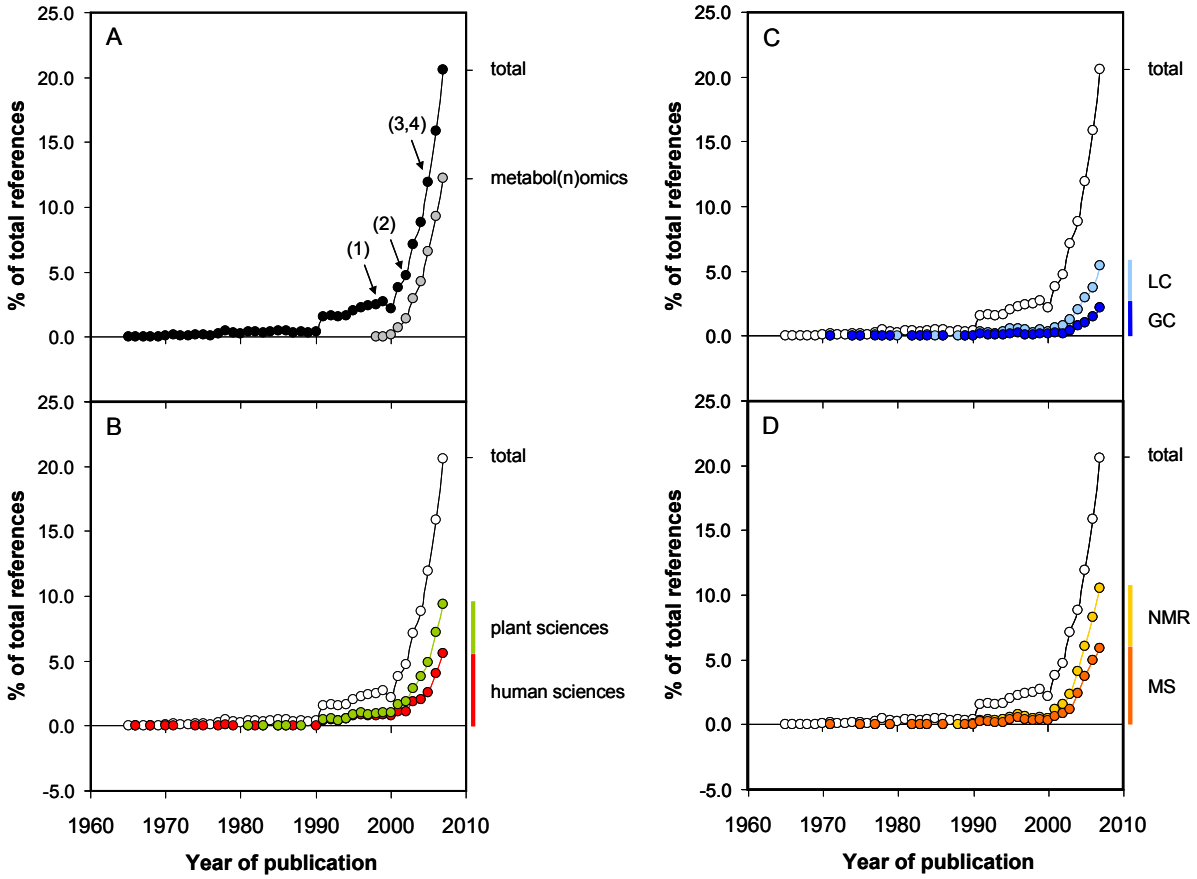


Figure 1 *Bibliographic concept analysis of published journal articles by year of publication as retrieved from the ISI citation index database.* A Boolean search string for the concepts metabolomics, metabonomics, metabolome, metabonome, and the exact terms "metabolite profiling", "metabolite profile", "metabolic profiling", "metabolic profile" and "fingerprinting or footprinting", respectively, with wildcards for alternative suffix usage returned 5,634 publications until December 31, 2007. A cumulative plot was chosen. **(A)** Appearance of the metabol(n)omics concept with earliest references in 1998 (1), the 1st International Conference on Plant Metabolomics in 2002 (2), the first issue of the Metabolomics journal and the 1st Annual International Conference of the Metabolomics Society both in 2005 (3,4), **(B)** contribution of plant and human sciences to the field, **(C)** use of gas and liquid chromatography and **(D)** use of mass spectrometry (MS) and nuclear magnetic resonance (NMR) technologies.

The objective of this thesis is the enhancement of one key metabolomic technology, namely, GC-MS based metabolite profiling, exploiting the complete available data resource and the full technological potential for routine application in molecular physiology and functional genomics. This work, which was initiated in September 1997 and has been resumed in 2001 after my commitment as a founding member of the Metanomics GmbH, contributes one essential building block towards truly comprehensive metabolic analyses and intends to lay the ground for the current transition to systems oriented, integrative investigations.

As I was involved in the emergence of this new field, I count my contributions to the paradigm shift underlying the metabolomic concept as a genuine and central achievement of this thesis. I will, therefore, first focus on the establishment of GC-MS profiling technology as an essential analytical technology to the novel metabolomic and systems biology tool box (cf. Kopka et al. 2004, Kopka 2006a, Steinhauser and Kopka 2007). After the description of these initial achievements, the key challenges of GC-MS based metabolite profiling and the resulting advanced aims of my thesis will be presented (cf. Kopka 2006b). Then, the respective advanced achievements of my research topics will be reported with an emphasis on the ongoing, yet unpublished technological and applied developments. Finally, the potential and the future perspectives of my work will be discussed at the end of each result paragraph rather than in a general discussion section.

RESEARCH TOPICS

GC-MS Based Metabolite Profiling in a Nutshell: Technological Criteria of Efficient Applications in Routine Metabolome Analysis

GC-MS technology has been used for decades in studies which aim at the exact quantification of metabolite pool size or metabolite flux and as a rule target single or small sets of metabolites. Today GC-MS is one of the most widely applied technology platforms in modern metabolomic studies (Fig. 1) as a rule covering more than 100 identified metabolites. However, early applications then called metabolite or metabolic profiling precede the modern metabolomic era, for example, applications to unravel herbicide mode of action (e.g. Sauter et al. 1988). Thus, multi-parallel GC-MS technology has experienced a renaissance brought about by the post-genomic requirements for high-throughput fingerprinting and metabolite profiling. Especially molecular plant physiology has embraced this new technology and a broad range of ecotypes or natural genetic variation (**Fiehn et al. 2000a [1]**, Hannah et al.

2006), breeding progeny (e.g. Schauer et al. 2006, Meyer et al. 2007), genetically modified genotypes (e.g. **Fiehn et al. 2000a [1]**, Roessner et al. 2001a, Roessner et al. 2001b, Roessner et al. 2002) and environmentally challenged plant systems (e.g. Cook et al. 2004; **Kaplan et al. 2004 [17]**, **Kaplan et al. 2007 [18]**, **Sanchez et al. 2008a [22]**, **Sanchez et al. 2008b [21]**) have been investigated. Metabolic phenotyping and analysis of respective phenocopies, defined as the full or partial similarity of metabolite profiles from different conditions, has become an integral part of plant functional genomics turning towards systems analysis (**Fiehn et al. 2000a [1]**, Roessner et al. 2002, Fernie et al. 2004, **Sanchez et al. 2008a [22]**), Eisenhut et al. 2008).

The analytical workflow of GC-MS analysis comprises seven essential steps, (i) extraction of a metabolite fraction from the biological sample, (ii) subsequent chemical derivatization, (iii) sample injection, (iv) gas chromatography, (v) ionization, (vi) mass detection, and most importantly (vii) automated chromatography data acquisition and processing (Kopka 2006a). All these aspects were considered when GC-MS was initially chosen in 1997-1998 as a metabolomic tool. The resulting operating procedure of the GC-MS metabolite profiling method was patented through the Metanomics GmbH (Herold et al. 2003a, Herold et al. 2003b, Herold et al. 2006). For application in academia a standard operating procedure (**Erban et al. 2007 [4]**) and a protocol of a subsequent method variation (**Lisec et al. 2006 [3]**) were published in the following years. Both procedures are in agreement with the recommendations of the Metabolomics Standards Initiative (MSI; e.g. Castle et al. 2006, Fiehn et al. 2007, Sumner et al. 2007). The next paragraphs will summarize and discuss the essential aspects of method establishment.

Metabolic Inactivation and Extraction. The initially targeted central metabolism but also the potential of future extended metabolite coverage were the first and ultimate criteria guiding method development. A genome-wide screening of metabolism using a complete single-gene knock-out population of *Arabidopsis thaliana* and populations of the same plant species resulting from single-gene over-expression of essentially every gene from *Escherichia coli* and *Saccharomyces cerevisiae* was intended and later completed by the Metanomics GmbH (e.g. Fernie et al. 2004). With the focus set to primary metabolism, mono-, di- and tri-saccharides, amino acids, organic acids and stable phosphorylated intermediates, were the targeted metabolite classes. The ultimate scope, however, was the full metabolome. Therefore, a non selective two step extraction procedure was established with a broad polarity range of solvents, comprising methanol:water and subsequent methanol:chloroform extractions of the biological sample. Metabolic inactivation of the biological sample was

performed by shock-freezing in liquid nitrogen with or without subsequent lyophilization. Metabolic inertness was maintained in subsequent steps by the enzyme inactivating properties of the employed organic solvents. The combined initial extracts were then separated by liquid partitioning into polar and lipophilic fractions which were finally dried for processing or storage. This basic protocol combined (i) effective metabolic inactivation of the sample and (ii) broad coverage of metabolites, essentially excluding only the macromolecular and volatile fractions and inevitably loosing labile metabolites (**Fiehn et al. 2000a [1]**, **Roessner et al. 2000 [2]**). Moreover, Fiehn et al. (2000a) **[1]** demonstrated the versatility of the GC-MS technology to establish multiple alternative profiles of the same sample, in this study complementing the information on polar metabolites with information on lipid composition after transmethylation. This first binary sample analysis allowed concentration of the lipids independently of the polar metabolites and thus detection of components, e.g. rare fatty acids as methyl esters, fatty alcohols or steroids, which would otherwise fall below the detection limits of a joined analysis. Such joined analyses were shown to be feasible using direct thermal desorption techniques, when fast scanning GC-time of flight (TOF)-MS instrumentation became available (e.g. Fiehn and Kind 2007). Other laboratories added the volatile fraction to the profiling toolbox using headspace solid phase micro-extraction (SPME) hyphenated to GC-MS (e.g. Tikunov et al. 2005, Tikunov et al. 2007). Furthermore, Weckwerth and co-authors (2004a) demonstrated that metabolome, transcriptome and proteome analyses of a single sample can be directly coupled. But the current state-of-the-art for integrative analysis of diverse systems levels is the generation of a deep frozen or lyophilized homogenate. This central stock sample is best split into aliquots for dedicated extraction methods which can then be perfectly optimized to the respective required amount and chemical nature of each targeted chemical fraction.

Chemical Derivatization. Targeted enzyme assays and high performance liquid chromatography (HPLC) methods were routine choices prior to the metabolomics era, only GC-MS coupled to a non-specific chemical derivatization, namely, trimethylsilylation, promised (Sauter et al. 1988) and later - after optimization to a two step procedure including methoxyamination in the presence of pyridine and trimethylsilylation - proved broad coverage of central metabolism within a single analysis of complex metabolic extracts (**Roessner et al. 2000 [2]**, **Fiehn et al. 2000a [1]**, Steinhauser and Kopka 2007). This two step derivatization has subsequently been tested and confirmed by numerous other scientists (e.g. Barsch et al. 2004, Broeckling et al. 2005, Gullberg et al. 2004, Sinha et al. 2004a, Sinha et al. 2004b,

Strelkov et al. 2004, O'Hagan et al. 2005). Today the method is applied in almost all biological sciences serving microbial, plant, animal and even human or clinical studies.

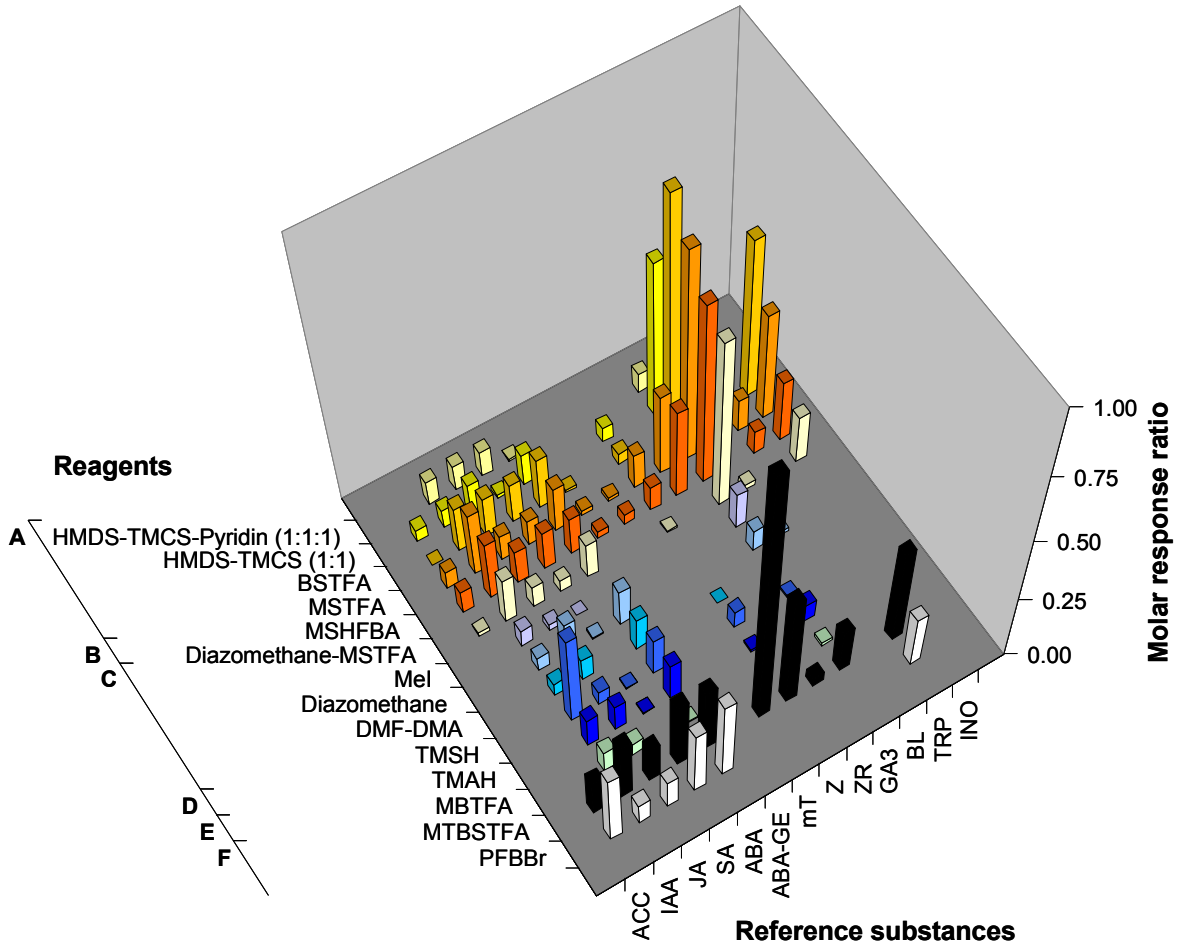


Figure 2 *Molar response ratios of the main derivative of representative phytohormones and reference substances demonstrate the comprehensive potential of trimethylsilylation.* Molar response ratios were calculated from electron impact GC-MS total ion currents of ~0.083 mg substance by normalization to the signal of an equal amount of 5α-cholestane used as an internal standard of each preparation. Derivatization experiments were performed with aliquots of the same reference mixture (n = 3). Reagents catalyzed **(A)** trimethylsilylations, **(B)** combined trimethylsilylation and methylation, **(C)** methylations, **(D)** trifluoroacetylation, **(E)** *tert.*-butyldimethylsilylation, and **(F)** pentafluorobenzylation. Common abbreviations of reagent names are given (cf. **Birkemeyer et al. 2003 [5]** for reaction details). Reference substances were: ABA, (6)-Abscisic acid, ABA-GE; (6)-abscisic acid-β-D-glucopyranosyl ester; ACC, 1-aminocyclopropane-1-carboxylic acid; BL, 24-epibrassinolide; GA3, gibberellic acid A3; IAA, indole-3-acetic acid; INO, *myo*-inositol; JA, (6)-jasmonic acid; mT, *meta*-topolin; SA, salicylic acid; Trp, DL-tryptophan; Z, *trans*-zeatin; ZR, *trans*-zeatin riboside. (Figure adapted from Table 1 of **Birkemeyer et al. 2003 [5]**).

So far no alternative derivatization scheme has replaced or complemented the initial method. The GC-MS technology, however, offers a large and diversified set of reagents which chemically modify a broad range of moieties or target specific chemical groups. Thus defined sets of non-volatile metabolites from complex extracts can be made amenable to gas chromatographic analysis. Besides alkoxyamination reactions which specifically target and stabilize carbonyl moieties, broad range alkylation- and acylation-reagents are available and compete with multiple silylation-reagents. These silylation reagents efficiently substitute acidic protons with trimethylsilyl- or *tert.*-butyldimethylsilyl-moieties. Most of the above reagents are in frequent use for compound-targeted GC-MS analyses, but have not yet been employed in large scale metabolomic studies (e.g. Blau and Halket 1993, Knapp 1979).

Birkemeyer at al. (2003) **[5]** using representative phytohormones and other reference metabolites covering a broad range of typical chemical properties demonstrated that the trimethylsilyl-reagents, N-methyl-N-(trimethylsilyl)-heptafluorobutyramide (MSHFBA) and N-methyl-N-(trimethylsilyl)-trifluoroacetamide (MSTFA) are indeed the most comprehensive chemical reagents and exhibit in comparison to other reagents sufficient sensitivity for efficient GC-MS profiling analysis (Fig. 2). Therefore, this thesis has so far been focused on the analysis of trimethylsilylated and methoxyaminated metabolites. However, two reagents showed the future potential for a more robust and sensitive but also more specific derivatization protocol, namely, N-methyl-N-(*tert.*-butyldimethylsilyl)-trifluoroacetamide (MTBSTFA) and pentafluoro-benzylbromide (PFBBr). Currently, only few examples of MTBSTFA based profiling exist (e.g. Jacobs et al. 2007), because MTBSTFA and PFBBr generate high molecular weight products and are prone to sterical hindrance effects. These properties restrict profiling applications to a smaller molecular weight range of metabolites and specifically exclude vicinal polyols (Fig. 2), which are typical moieties of sugars, sugar alcohols and sugar conjugates, metabolite classes that dominate conventional comprehensive metabolic extracts. With these limitations the best application of MTBSTFA and PFBBr reagents will be in combination with fractionation or pre-purification steps which exclude sugars and other high molecular weight metabolites. Such fractions will allow enrichment of those trace metabolites which currently fall below the detection limit of conventional GC-MS profiling. A low through-put application targeting multiple phytohormones was established by Birkemeyer and co-authors (**2003**) **[5]** but was not suitable for large scale screening. Nevertheless, this reagent tool box may become increasingly important for future high-throughput applications once the initial bottle necks of automated chromatography data processing and compound identification have been solved.

Sample Injection and Chromatography. Gas chromatography is already a highly automated analytical technology suitable for the processing and high-resolution separation of small and complex samples. As such GC-MS was ready to be employed for metabolome analyses and only the choices of automated injection procedure and selection of capillary column had to be made for method establishment.

GC features liquid and volatile injection modes. Volatile injection is typically used without chemical derivatization. Respective methods, namely, headspace, cold trapping and SPME-GC-MS, have been used to profile the volatile metabolic complement (e.g. Tikunov et al. 2005, Tikunov et al. 2007). Volatile injection modes have also been applied to trimethylsilylated metabolite preparations using a process named vapor phase extraction (VPE). VPE was shown to operate with a range of commonly used reagents and has the potential to represent a robust technique mostly because non-derivatized material is prevented from contaminating the GC-MS system (Schmelz et al. 2003, Schmelz et al. 2004). But VPE does not allow high-throughput sample processing. In view of the high potential for automation, routine liquid injections using split or splitless modes were the methods of choice (e.g. **Fiehn et al. 2000a [1]**, **Roessner et al. 2000 [2]**, **Lisec et al. 2006 [3]**, **Erban et al. 2007 [4]**). The splitless mode is most widely applied, because the complete derivatized liquid sample, typically 0.5-2.0 µl, is transferred to GC-MS analysis and matrix effects discriminating low versus high boiling chemical derivatives are avoided. In contrast, split injection typically transfers only 1/10-1/100 of the sample onto the GC column. Thus, using the splitless mode the required amount of biological sample can also be minimized typically to 1-100 mg fresh weight (FW) and down to a final reagent volume of 30 µl for robust robotized chemical derivatization (**Erban et al. 2007 [4]**). Simultaneously Erban and co-authors (**2007**) **[4]** implemented automated in-line derivatization of dried extracts coupled to exactly timed injection after derivatization for routine GC-MS analyses. These developments served to counteract the "time-on-the-tray effect" caused by analyte decay of partially labile derivatized compounds.

The choice of capillary column was motivated by two and, as turned out later, conflicting aspects: the optimum separation of complex mixtures and the stability of the capillary to extreme temperature ramping protocols. Extreme operating temperatures are necessary for best coverage of the extreme volatility range of metabolite derivatives. The basic choice was a fused silica capillary column with a composite phenyl-dimethylpolysiloxane stationary phase and typical dimensions, namely, 0.25 µm film thickness, 30 m length and 0.25 mm inner diameter. Such stationary phases are ideally

compatible with silylation reagents, because free hydroxyl-groups which are generated throughout the aging process of polysiloxane phases are continuously inactivated. First implementations were performed with 50%-phenyl-dimethylpolysiloxane phases for best separation of typical biological isomers from central metabolism, for example citrate and isocitrate or abundant sugars, such as sucrose, glucose, galactose, mannose and fructose (e.g. **Roessner et al. 2000 [2]**). However, operating time of these columns was restricted to a few hundred analyses. Subsequent optimization led to the use of integrated guard columns and arylene stabilized equivalents of 5%-phenyl-dimethylpolysiloxane phases (e.g. **Erban et al. 2007 [4]**) or 35%-phenyl-dimethylpolysiloxane phases (e.g. **Lisec et al. 2006 [3]**), respectively. Both types can be operated at high temperatures, 350°C and 330°C, respectively. Under typical operating conditions these columns allow >1000 analyses with only minor changes of chromatographic retention behaviour (**Strehmel et al. 2008 [14]**). The 5%-phenyl-dimethylpolysiloxane phases have been most widely applied, with variants ranging from fast GC of 15 min analysis time (Gullberg et al. 2004) to slow but high-resolution chromatography of more than 70 min duration (Barsch et al. 2004). The resulting compendium of method variations has been highly valuable for optimizing data exchange between laboratories and for comparison of GC-MS performance (e.g. **Schauer et al. 2005 [10]**, **Strehmel et al. 2008 [14]**). Novel hyphenations of two dimensional GC (GCxGC) to mass spectrometry now combine arylene type 5%-phenyl-dimethylpolysiloxane stationary phases for 1st dimension separation to short narrow bore capillaries of 50%-phenyl-dimethylpolysiloxane phases for the 2nd dimension (Fig. 3) and robustness of analysis still has to be explored.

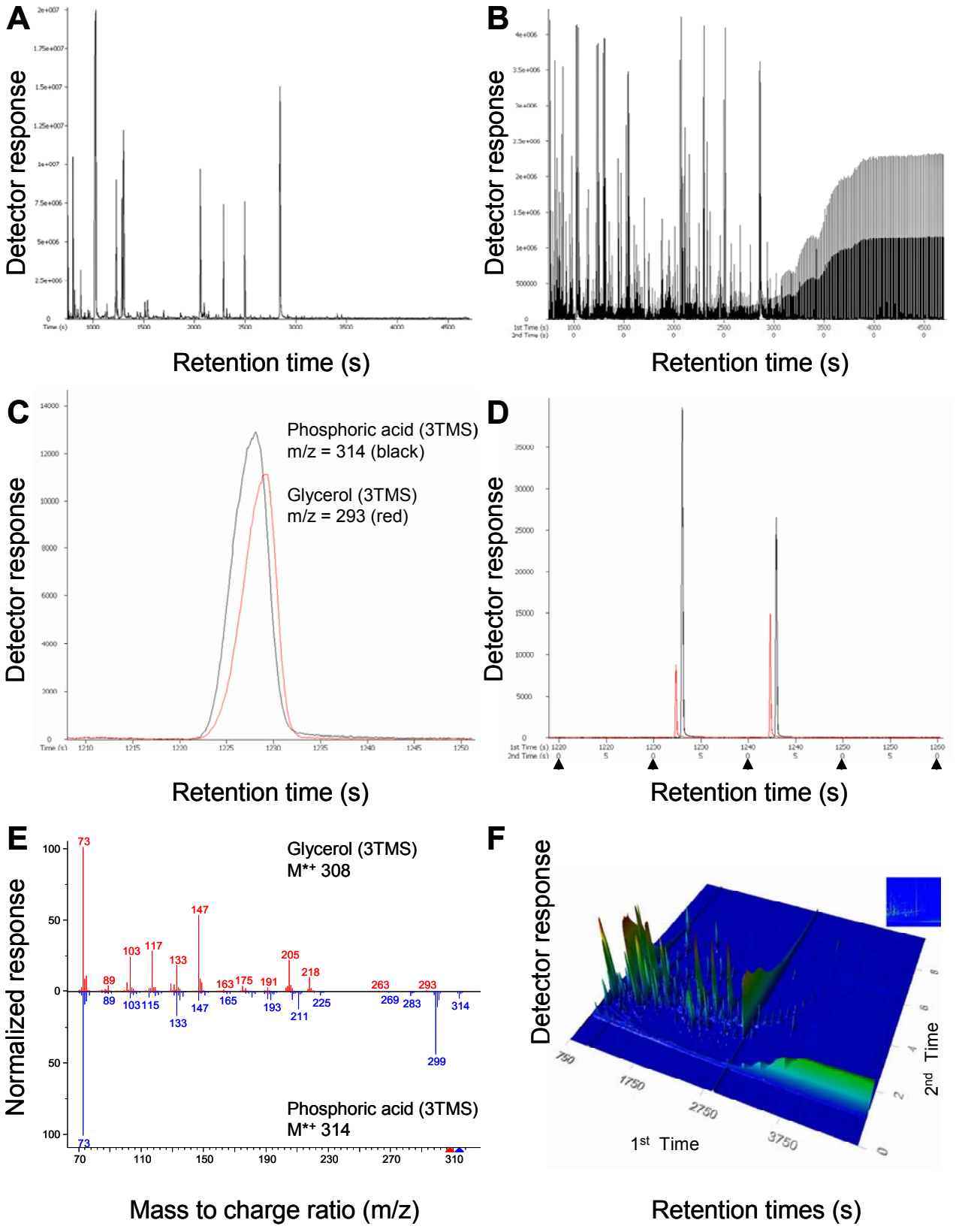
Ionization and Mass Spectrometry. Compounds are ionized prior to mass detection as they elute from the GC. Electron impact (EI) ionization is most widely used, as it is the technology which is least susceptible to ionization suppression effects and produces characteristic fragmentation patterns with highly reproducible quantitative ratios. Moreover, these EI fragmentations patterns, commonly called mass spectra, are comparable when recorded by several mass spectrometric technologies. The drawback of EI is the high degree of fragmentation resulting in low abundance or even absence of the non-fragmented molecular ion. For this reason chemical ionization (CI) methods should be suitable for retaining non-fragmented molecular ions and may in theory result in a gain of sensitivity and selectivity. However, positive and negative CI modes did not emerge in high-throughput applications as chemical ionization appears to be less robust and efficient compared to EI.

One of the most important underlying reasons for the success of GC-MS in chemical and metabolic analysis may be the optimum fit of the low amounts of substance required for GC separation and the high, inherent sensitivity of mass spectrometric detection. For example, the low sensitivity of nuclear magnetic resonance (NMR) technology, the prevailing alternative of metabolite detection in the metabolomic field (Fig. 1), has so far impaired successful in-line hyphenation to GC. Indeed an efficient GC-NMR coupling would represent a paradigm change for metabolic flux and pool size analysis. Today GC can be efficiently hyphenated to diverse mass-detection devices, including sector field detectors, quadrupole detectors (QUAD), ion trap technology (ITP), and time-of-flight detectors (TOF). The choice of detectors depends on the intended analytical niche. GC-MS systems with QUAD detection are the most widely spread for routine analysis. ITP technology allows two- or multi-dimensional mass spectrometric analysis (MS^n) for structural elucidation and multi-targeted quantification of trace compounds (e.g. Mueller et al. 2002, **Birkemeyer et al. 2003 [5]**). TOF detection can either be tuned to fast scanning rates but low mass resolution, typically 0.1-1.0 amu (e.g. Dalluge et al. 2002a, Dalluge et al. 2002b, van Deursen et al. 2000, Veriotti et al. 2000, Veriotti et al. 2001, Vreuls et al. 1999), or to high mass precision comparable to sector field systems but lower scanning rates. High mass precision allows good prediction of molecular masses for structural elucidation, but fast scanning TOF has greatly enhanced metabolic profiling applications by enabling good mass spectral recording and deconvolution of co-eluting compounds. From 1997-1998 before GC-TOF-MS systems became commercially available until 2003 this thesis contributed to the successful application of QUAD technology in metabolome analysis (**Fiehn et al. 2000a [1]**; Fiehn et al. 2000b, **Roessner et al. 2000 [2]**, Roessner et al. 2001a, Roessner et al. 2001b). Then the transition to GC-TOF-MS based metabolite profiling was made (Fiehn and Weckwerth 2003, **Wagner et al. 2003 [8]**). Currently, fast scanning GC-TOF-MS may further revolutionize metabolome analysis. GC-TOF-MS enables the most advanced commercial instruments in the GC-MS field, namely, two dimensional GCxGC-TOF-MS systems (Marriot et al. 2000, Shellie et al. 2001, Marriott and Shellie 2002, Ryan et al. 2004; Sinha et al. 2004a, Sinha et al. 2004b, Sinha et al. 2004c) which use 2-10 s secondary GC separations and are, therefore, dependent of fast scanning mass detectors.

While the data structure of GCxGC-TOF-MS is superior to conventional GC-TOF-MS (Fig. 3), the huge size of the acquired chromatography data files and the lack of software support for most aspects of automated medium to high-throughput chromatography data processing currently impair routine metabolite fingerprinting or profiling appliations.

Figure 3 (>) *GC-TOF-MS-data structure compared to GCxGC-TOF-MS using identical extracts of consumable rice seeds.* Total ion current chromatograms of a GC-TOF-MS system, 20 scans s^{-1} **(A)**, are compared to GCxGC-TOF-MS, 100 scans s^{-1} **(B)**, analysis. Zooming in on selective ions **(C-D)** exemplifies the use of selective mass fragments, e.g. m/z 293 representing per-trimethylsilylated glycerol (3 TMS) and m/z 314 representing co-eluting phosphoric acid (3TMS), respectively. GC-TOF-MS **(C)** utilizes selective mass fragments of these co-eluting compounds for quantification; in contrast GCxGC-TOF-MS **(D)** employing secondary chromatography with altered elution sequence of compounds exhibits enhanced selectivity by baseline separation of otherwise co-eluting compounds. **(E)** As a result the pure mass spectra of these compounds can be obtained after GCxGC-TOF-MS without deconvolution. Baseline subtracted mass spectra comprise both, the common and the specific mass fragments of each compound. The use of 10 s **(D, arrows)** or shorter secondary chromatograms (modulations) generates multiple peaks of the same compound which are distributed among subsequent modulations. For quantification and visualization **(F)** the GCxGC information of single compounds must be reconstituted from subsequent modulations. Note the enhanced apparent sensitivity of GCxGC-TOF-MS **(B)** which is caused by inherent cryo-sampling and resulting substance enrichment and peak sharpening (METAPHOR EU-FOOD-CT-2006-036220 project, unpublished).

Chromatography Data Acquisition and Processing. GC-MS analyses generate an enormous amount of primary data, which depend on choices of mass range, chromatographic duration and mass spectral scanning rate (Fig. 3). Typical GC-TOF-MS files of ~66 min duration acquired with 20 mass spectral scans s^{-1} and the mass range set to 50-1,000 amu, may amount to almost 80,000 full mass spectra and 400-430 Mb file size. In comparison, a GCxGC-TOF-MS file of identical duration and mass range will require at least 100 scans s^{-1} and, thus, typically generates almost 400,000 primary mass spectral scans deposited into 2,000-2,100 Mb files. Instrument manufacturers provide a standard file interchange format for GC-MS data, called NetCDF, and more or less efficient software solutions for data processing, which are typically tuned to metabolite targeted quantification. Most of the vendor software tools are not applicable for the non-targeted and - within technological limitations - non-biased metabolomic approaches. Therefore, the lack of software solutions for metabolic fingerprinting and profiling was the dominant impediment of modern metabolomics and may have led to the abortion of previous implementation attempts in the "pre-genomic" phase (e.g. Jellum et al. 1975, Jellum 1977, Jellum 1979).


A
Detector response
Retention time (s)
B
Detector response
Retention time (s)
C
Detector response
Phosphoric acid (3TMS)
m/z = 314 (black)
Glycerol (3TMS)
m/z = 293 (red)
Retention time (s)
D
Detector response
Retention time (s)
E
Normalized response
Glycerol (3TMS)
M*+ 308
Phosphoric acid (3TMS)
M*+ 314
Mass to charge ratio (m/z)
F
Detector response
1st Time
2nd Time
Retention times (s)

In the post-genomic phase GC-MS had a head-start of software automation compared to other mass spectral technologies. Specifically tools for automated mass spectral deconvolution and mass spectral matching to large custom or commercial spectrum libraries were already in place. Automated deconvolution may be seen as one of the principal pre-existing achievements contributing to the success of modern GC-MS profiling analyses. Deconvolution is defined as automated mathematical decomposition of the multiple partially co-eluting mass spectra which constitute a GC-MS data file (e.g. Fig. 3C). The most widely applied software solutions for this process are the open access Automated Mass Spectral Deconvolution and Identification Software, AMDIS (Halket et al. 1999; Stein 1999), the commercial vendor-independent AnalyzerPro suite, http://www.spectralworks.com, and the vendor specific ChromaTOF software for automated processing of GC- and GCxGC-TOF-MS data from Pegasus systems (LECO, St. Joseph, MI, USA). These deconvolution tools but also other software solutions (e.g. Shao et al. 2004) perform (i) mass resolved baseline subtraction of electronic and chemical noise, (ii) assignment of retention times and/or retention time indices (RI) to chromatographic peak apices or deconvoluted mass spectra and (iii) deconvolution of mass spectra from closely co-eluting compounds. A recent comparative study found ChromaTOF (v2.15) software to be superior compared to AMDIS and AnalyzerPro, but still the deconvolution process was found to be error prone (Lu et al. 2008).

The availability of commercial mass spectral libraries and a commonly accepted standard for mass spectral matching (Ausloos et al. 1999; Stein 1999) may have been the second equally important initial boost for the success of GC-MS based metabolite profiling. The GC-MS technology is in this respect more advanced than LC-MS (Halket et al. 2005). A common standard, namely, the mass spectral search and comparison software of the National Institute of Standards and Technology (NIST, Gaithersburg, MD, USA) has been integrated into the customized operating software of most GC-MS manufacturers. Commercial mass spectral libraries, as provided by NIST or Wiley publishers, and customized user specific collections (e.g. **Kopka et al. 2005 [11]**, **Schauer et al. 2005 [10]**) can be used in combination. While large scale retention index libraries were not available until 2005 the new versions, NIST05 and NIST08, provide RI information. But RI information is not integrated into the automated matching protocols.

Technical Performance. All aspects mentioned above contributed to the overall high reproducibility and robustness of GC-MS based metabolite profiling with the technological standard deviations typically lower than the biological variation under controlled experimental conditions. Exemplary analyses of detection limits, linear range of

quantification, usually 2-3 orders of magnitude, and recovery, typically 70-130 %, lead to routine applications which allowed large scale sample screening by relative quantification of detector signals. Relative quantification still routinely performed by comparison to standardized biological reference samples or more generally applicable chemically defined mixtures of authenticated reference substances (e.g. **Fiehn et al. 2000a [1]**, **Roessner et al. 2000 [2]**, Herold et al. 2003a, Herold et al. 2003b, Herold et al. 2006, **Strehmel et al. 2008 [14]**). Typically at least 300, sometimes up to ~1,000, chemical components are resolved. Of these compounds about 100-150 may represent internal standards or reagent and laboratory contaminations which need to be excluded from subsequent analyses (**Wagner et al. 2003 [8]**). The resulting set of useful data depends on the nature of the biological sample and typically comprises ~50-150 metabolites which can be recognized and accessed for profiling experiments and a large portion of yet non-identified compounds for fingerprinting purposes.

In conclusion, non-targeted GC-MS based profiling of methoxyaminated and trimethylsilylated comprehensive metabolic extracts has been implemented and brought to routine analysis in the course of this thesis. The method has since been continuously enhanced and refined. The accepted role of GC-MS based metabolite profiling in the metabolomics field may be defined as monitoring of primary metabolite patterns from central metabolism in combination with facultative co-analysis, screening and discovery of yet non-identified or unexpected small secondary metabolites which fall into the analytical scope of this tool. The method has been proven to be suitable for quantitative purposes and has become highly versatile beyond the initial applications in plant physiology. The success was based on an ideal combination of (i) comparatively low investment and operating costs, (ii) bench-top size and highly automated instrumentation, (iii) superior broad non-selective derivatization of metabolites, (iv) stable high-resolution gas chromatographic separation of complex mixtures, (v) selective quantitative detection of co-eluting compounds by specific mass fragments and (vi) highly reproducible mass spectral fragmentation with basic software tools and libraries for manually supervised mass spectral analysis and matching in place.

For these reasons GC-MS based metabolite profiling has by some authors been judged to represent the “gold standard” of metabolite profiling (e.g. Harrigan and Goodacre 2003, Lu et al. 2008). The downsides or limitations of GC-MS profiling may be shortly summarized in the by the following criteria: The range of accessible metabolites is restricted to small, volatile, and non-thermo-labile metabolites. For most applications chemical derivatization is required. This necessity leads to chemical modifications and loss of reactive or instable metabolites. Other than NMR, GC-MS is less easily amenable to exact quantification, and

requires for this purpose metabolite specific testing of (i) detection limits, (ii) recovery, (iii) linearity of detector response and (iv) internal as well as external quantitative calibration. Finally, the advantage of small sample size and high mass spectral fragmentation which both contribute to the efficient multi-parallel top-down identification of GC-MS components, facilitating the recognition of known compounds by comparison to authenticated reference substances, interferes with the bottom-up identification of yet unknown compounds. This identification procedure requires pure compounds. Preparation of sufficiently high amounts after GC separation is a highly difficult task. These limitations define the key challenges of current GC-MS based metabolite profiling and have been addressed by this thesis as will be reported in the following.

Key Challenges and Technological Aims of Enhanced GC-MS Based Metabolite Profiling

Routine GC-MS based metabolite profiling has provided a combination of technological features in time for post-genomic applications. These properties have led to a rapid and continued succession of successful applications within almost all biosciences. The inherent versatility of GC-MS instrumentation now culminates in contributions to systems biology, bridging both extremes of systems biology, on the one side large scale metabolic screenings and, on the other side, combination of flux and pool size quantification for the mathematical modeling of metabolism. A set of key challenges became apparent already when the first profiling experiments were performed but also in the subsequent process of establishing automated routine workflows which lead from sample processing and to the final physiological interpretation of profiling results. Starting with the achievements of the basic metabolite profiling concept, reported above, this thesis tackles a key metabolomic challenge, i.e. to enhance GC-MS technology towards an efficient analysis of molecular metabolic processes in the new era of systems biology. I have chosen the following projects because of their essential contributions to this ultimate aim.

Project 1: Development of a Mass Spectral and Retention Index Reference Library. The perhaps most astonishing discovery of GC-MS based metabolite profiling was the large number of metabolites, which were observable in the highly complex extracts of plants (**Fiehn et al. 2000a [1]**, Fiehn et al. 2000b), and later in the profiles of many other species. A large diversity of primary metabolites was immediately identified by mass spectral matching to commercial spectral libraries. These matches were validated by the proof of compound

identity using standard addition experiments of commercially available and authenticated chemical reference substances. However, compared to the high complexity only less than 25% of the observed compounds was identified, using what we now call the top-down identification process. In consequence, the fundamental aims of this thesis are the establishment, enhancement and dissemination of a reference library, tuned to the needs of GC-MS profiling and harboring chemical information on both, pure reference substances and yet non-identified metabolic components of defined biological material.

Project 2: Automated Data Processing of Complex GC-MS Based Metabolite Profiles. The basic analytical technology was readily automated and standardized following good laboratory practice (**Lisec et al. 2006 [3]**, **Erban et al. 2007 [4]**). The first implementation of chromatography data processing was a slow, manual and expert user dependent process. The lack of standardized and automated chromatography data processing represented the first bottle neck for reproducible high-throughput analyses of profiling experiments which rapidly increased in size from ~20 to even more than 1,000 GC-MS chromatograms constituting a single experiment. The required key processing functions were (i) the unambiguous recognition of metabolites, (ii) the proper choice of the optimum mass feature(s) for relative or absolute quantification and (iii) - prerequisite of these functions - the global, non-biased data pre-processing generating a numerical data matrix with access to all observed mass features of an experimental data set. This last feature covers three additional demands of metabolomic studies, namely, the non-targeted discovery of novel and unexpected metabolites, the option to implement internal standardization by stable isotope labeled substances and, finally, the ultimate goal to add flux analysis to the tool box by comprehensive monitoring of mass isotopomer distributions. Therefore, the second fundamental aim of this thesis is the development of a software tool for the purpose of standardized GC-MS data pre-processing including a comprehensive data matrix generation and the support for automated metabolite recognition.

Project 3: The Golm Metabolome Database (GMD). Both, the high number of metabolic components and chemical reference data, as well as the large amount of numerical information generated by the standardized processing of each GC-MS based profiling experiment proved to be difficult to handle and to communicate between scientists. With standardized protocols and data formats developed in the first two projects the necessity to establish a custom database for archiving GC-MS based profiling experiments and reference data became obvious. For this purpose, the Golm Metabolome Database (GMD) was

initiated. This database is now being continuously extended to serve the growing needs of enhanced GC-MS based metabolite profiling.

Project 4: Enhanced Metabolite Profiling using Mass Isotopomer Ratios (ITR). Early publications proposed the use of stable isotope labeled internal standardization for enhanced quantitative accuracy and precision (e.g. **Fiehn et al. 2000a [1]**). As a multi-targeted technology GC-MS based metabolite profiling can obviously not be optimized to each of the chemically divers constituent metabolites but must aim for a global optimum. Therefore, the concept of synthetically labeled chemical standards has been extended in this thesis towards full *in vivo*-labeling of whole organisms and the application of such labeled and defined biological reference material as multiplexed quantitative internal standards. Thus, comprehensive standardization was expected to become achievable and to substantially reduce the technological variability caused by compound specific recovery and matrix effects. Ultimately, the improvement of the GC-MS profiling tool will justify added costs and experimental demands of ITR.

Project 5: Towards Combined Metabolite Pool Size and Flux Analysis. The causal interpretation of physiological changes in metabolite pool sizes is limited as metabolite pools are typically controlled by both, anabolic and catabolic reactions. Therefore, the observation of a change in pool size will in most cases have ambiguous interpretations, namely, the possibility of altered synthesis, altered degradation or a combination of both. Additional information on metabolic flux may substantially enhance the understanding of metabolic systems. In conclusion, the combination of information on metabolite concentration, metabolic flux and thermodynamic considerations as proposed by Kummel et al. (2006) should represent the best conceivable data resource for metabolic modeling. This expectation motivated the attempt to establish a method for the combined concentration and flux determination of soluble metabolite pools in biological systems.

Project 6: The Metabolic Component of Plant Environmental Stress Acclimation. Plant responses to environmental stresses typically comprise metabolic components which have previously been described by metabolite targeted studies. Even though many metabolic responses have been known for decades, only a few of these metabolic responses are functionally understood. For this reason the GC-MS based metabolite profiling technology was applied - initially in descriptive studies - tackling the comprehensive assessment of metabolic reprogramming which has been enabled by metabolite profiling. Thus, the elucidation and understanding of the metabolic responses to abiotic environmental stresses, such as temperature and salt, has become a focus of this thesis. Projects have been initiated in

co-operation with expert laboratories on stress physiology and are now driven by the metabolic knowledge base of my laboratory.

In the following paragraphs the above projects will be shortly discussed with a concluding summary of each key achievement and current state-of-the-art. A specific outlook will be given following each project description. These paragraphs will include trends and visions of future technological or physiological applications.

Development of a Mass Spectral and Retention Index Reference Library

Early manual inventories of methoxyaminated and trimethylsilylated metabolite fractions and later routine data mining of GC-MS based metabolite profiles led to striking observations of chemical complexity. After subtraction of chemical artifacts which may be caused by reagent- and solvent-contaminations (**Wagner et al. 2003 [8]**) a metabolic window of typically 300-500 analytes, i.e. chemical derivatives of metabolites, can be detected. Due to the enhanced resolution and sensitivity of enhanced GCxGC-TOF-MS instrumentation this number is expected to increase further by a factor of 2-10 (e.g. Nielsen and Oliver 2005) once this technology will have entered routine applications. The analytes of GC-MS profiling are characterized by mass spectrum and 1-dimensional (GC-TOF-MS) or 2-dimensional (GCxGC-TOF-MS) chromatographic retention. Chromatographic retention is typically standardized by *n*-alkanes spiked into GC experiments and is expressed as a retention index (RI). These combined chemo-physical properties, so-called mass spectral tags (MSTs). The MST concept was proposed and adopted during the early steps of mass spectral library generation and serves naming and archiving of the initially unknown components of GC-MS profiles (cf. **Desbrosses et al. 2005a [20]** refined by Kopka 2006b). In short, MSTs represent the means of analyte recognition and identification. In the following the term "identification" will be used for the first elucidation of the chemical structure represented by a MST, whereas the term "recognition" will be used for the subsequent matching process that is required in the workflow leading from non-targeted fingerprinting to metabolite focused profiling experiments.

Currently, the majority of MSTs is not chemically characterized and linked to metabolite structures. Indeed routine application of GC-MS profiling to biological samples, which have not been investigated before, or have been exposed to new chemical, biotic or abiotic stress conditions, or result from new natural or genetically engineered genotypes, still yield novel MSTs. These elicited MSTs may represent valuable diagnostic markers, which

are currently not amenable to functional analysis and physiological interpretation, because the chemical identity and thus metabolic pathway connectivity is not known. Therefore, the full potential of the metabolic profiling technique remains only partially accessed. The future of GC-MS based profiling will depend on the continued identification process of all its component analytes. Chemical identification towards this aim may follow two strategies, namely, the "top-down" and the "bottom-up" approaches (cf. Steinhauser and Kopka 2007) which are described in the following.

Top-down Identification. The top-down identification process of analytes uses authenticated reference compounds in standard addition experiments and performs identification by both, chromatographic retention and mass spectral fragmentation, where the mass spectral fragmentation of the MST must confirm the expected analyte structure after methoxyamination and trimethylsilylation of the underlying metabolite (cf. Fig. 4-5). Criteria for confirmation are expected retention indices, presence of molecular ions M^{*+}, and typical fragment ions such as $M\text{-}15^{+}$ or previously reported specific fragments and neutral losses of the respective compound classes. Commercial mass spectral libraries comprising more than 191,436 non-redundant compounds for mass spectral matching, e.g. the NIST/EPA/NIH mass spectral library version 2.0, June 25th 2008, and the NIST08 software (http://www.nist.gov/data/nist1a.htm), provide standardized tools for manually supervised identification. Because of the speed and rich resource of previous knowledge, top-down identification appears to be the most effective and least error prone strategy. It may fit ideally to the proposed standardization (Bino et al. 2004; Jenkins et al. 2004) and continued refining efforts of reporting standards for the chemical analysis in the metabolomics field (Sumner et al. 2007). The commercially available libraries were, however, not tuned to the biological samples of the metabolomic field and the initial success rate was only ~25% of analytes identified.

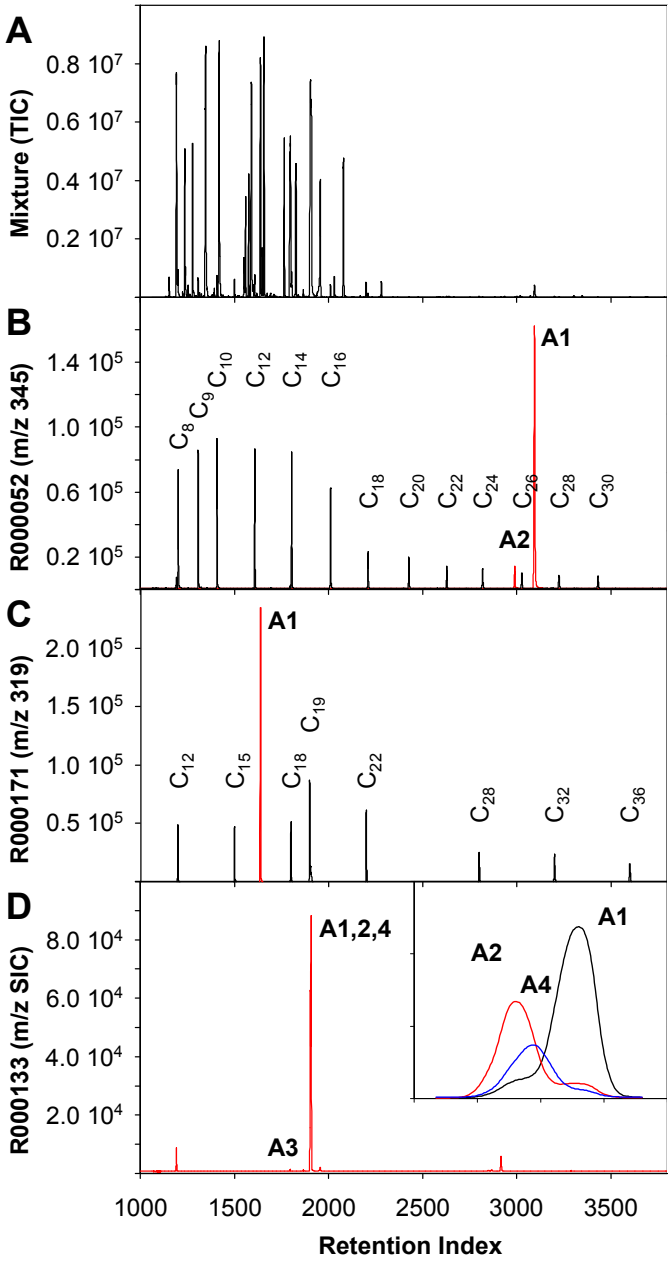


Figure 4 *Exemplary standard addition experiment of authenticated reference substances.* A mixture of 20-25 reference substances **(A)**, biological reference material (not shown) and pure substances **(B-D)** are analyzed in total ion chromatography (TIC) mode following either the method of Erban et al. (2007) **[4]**, shown here, or Lisec et al. (2006) **[3]**. Each sample contains both, a series of fatty acid methyl esters (**B**; C_8-C_{30} monitored by an extracted selective ion chromatogram, SIC, at m/z 87) and a series of *n*-alkanes (**C**; C_{12}-C_{36} monitored by m/z 85). These compounds are used for internal standardization of chromatographic retention and retention index calculation (cf. **Strehmel et al. 2008 [14]**). The van den Dool and Kratz (1963) alkane retention index is used here. Mixtures and biological reference material demonstrate co-elution behavior. Single reference substances, for example chlorogenic acid **(B)**, ribitol **(C)**, and gluconic acid **(D)** allow extraction of chemically derivatized analytes (A1-A4, red SICs) and impurities with respective MST information. Impurities may provide valuable additional metabolite identifications. For example, commercial chlorogenic acid (**B**; A1) contains a geometric Z-isomer (**B**; A2). Other preparations, for example commercial gluconic acid, contain multiple impurities (**C**; A1-A4) which may co-elute (cf. **C**; insert) and thus require automated mass spectral deconvolution, cf. Fig. 5 (Bölling C, Erban A, Kopka J and Willmitzer L, unpublished).

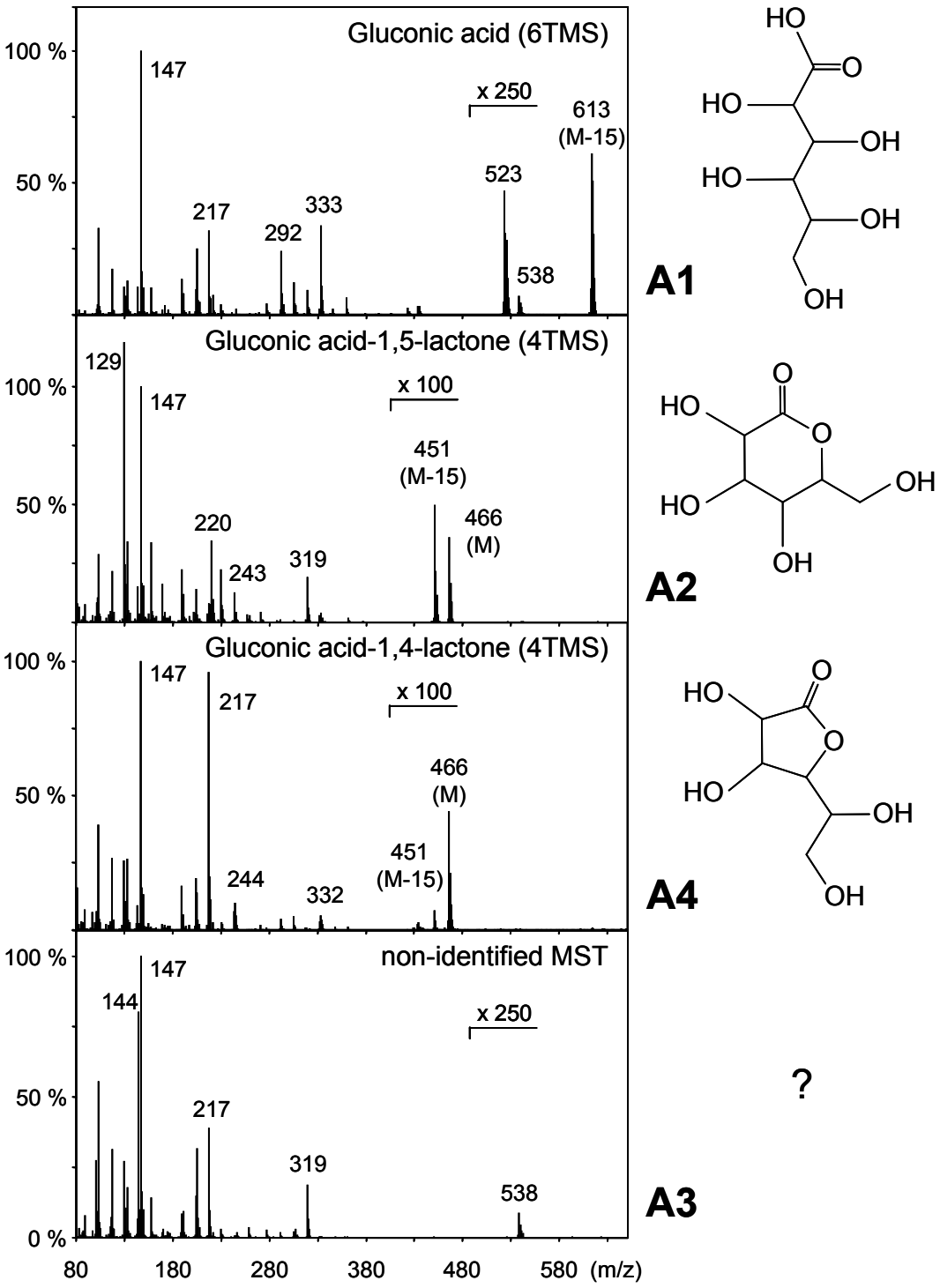


Figure 5 *Mass spectral analyses of four analytes detected by GC-TOF-MS profiling of a commercial gluconic acid preparation.* Analyte A1 was expected and confirmed by $M\text{-}15^{+}$ = 613 as well as best mass spectral similarity to the respective products of commercially available galactonic acid and gulonic acid. A2 and A4 represent two alternative lactones, M^{*+} = 466 and $M\text{-}15^{+}$ = 451, which are impurities formed in aqueous solution by loss of a water molecule from gluconic acid. These analytes were later confirmed by authenticated lactone preparations. A3 remained non-identified and was archived as an impurity MSTs of gluconic acid for future identification (Bölling C, Erban A, Kopka J and Willmitzer L, unpublished).

In the course of this thesis a cooperative international effort of collecting reference libraries from expert GC-MS profiling laboratories was initiated to serve effective top-down identification (**Schauer et al. 2005 [10]**). The resulting combined data were disseminated to the metabolomic field through the Golm Metabolome Database (GMD, http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html, **Kopka et al. 2005 [11]**). The initial basis of acquiring a reference substance in each of the participating laboratories was a mass spectral match of an observed MST to a mass spectral library, which was then experimentally confirmed by commercially available reference substances.

In extension of this basic approach a systematic analysis of metabolic pathways for compounds which may fall into the scope of GC-MS based metabolite profiling was initiated (Boelling C, Erban A, Willmitzer L and Kopka J, unpublished). Species specific and generalized metabolic networks are accessible to the academic domain (Bader et al. 2006), for example the Human Metabolome Project (Wishart et al. 2007, Wishart 2007), the BioCyc suite of databases (Karp et al. 2005) comprising MetaCyc (Caspi et al. 2006), EcoCyc dedicated to *Escherichia coli* K12 (Keseler et al. 2005) and AraCyc for the plant model organism (Zhang et al. 2005). In addition suitable tools such as BioPathAt (Lange and Ghassemian 2005), Mapman (Usadel et al. 2005), MetNet (Wurtele et al. 2003) and PaVESy (**Luedemann et al. 2004 [9]**), the later developed in the framework of this thesis, allow pathway visualization, manipulation and network analysis. Perhaps the most comprehensive resource of metabolic network reconstructions is the long-standing Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa 1997; Kanehisa and Goto 2000; Kanehisa et al. 2002, Kanehisa et al. 2006), which now also includes EST based information (Masoudi-Nejad et al. 2007). The KEGG pathway and metabolite annotations (Goto et al. 1998) comprise 14,549 metabolite entries, release 42.0, April 1, 2007, which are frequently cross-referenced to other large inventories of small molecules, such as PubChem (http://pubchem.ncbi.nlm.nih.gov/) or ChEBI (http://www.ebi.ac.uk/chebi/). The ongoing process of compound acquisition and analysis in my laboratory currently adds up to a retention index and mass spectral library of 1,632 reference substances and 2,437 non-redundant MSTs of respective analytes (Strehmel N, Hummel J and Kopka J, personal communication). These analytes cover diverse compound classes such as organic acids, amino acids, sugars, alcohols, polyols, aldehydes, amines, amides, imides, lactams, alkaloids, calystegines, fatty acids, alkanes, terpenoids, chalcones, stilbenes, flavonoids, indoles, purines, pyrimidines, nucleosides, nucleotides and diverse metabolically relevant conjugates. Depending on the biological object 30-40% of the observed analytes, which typically represent ~50-150 metabolites, are currently identified in

routine GC-MS profiling experiments and can be linked to respective known metabolite pathways of the analyzed biological object.

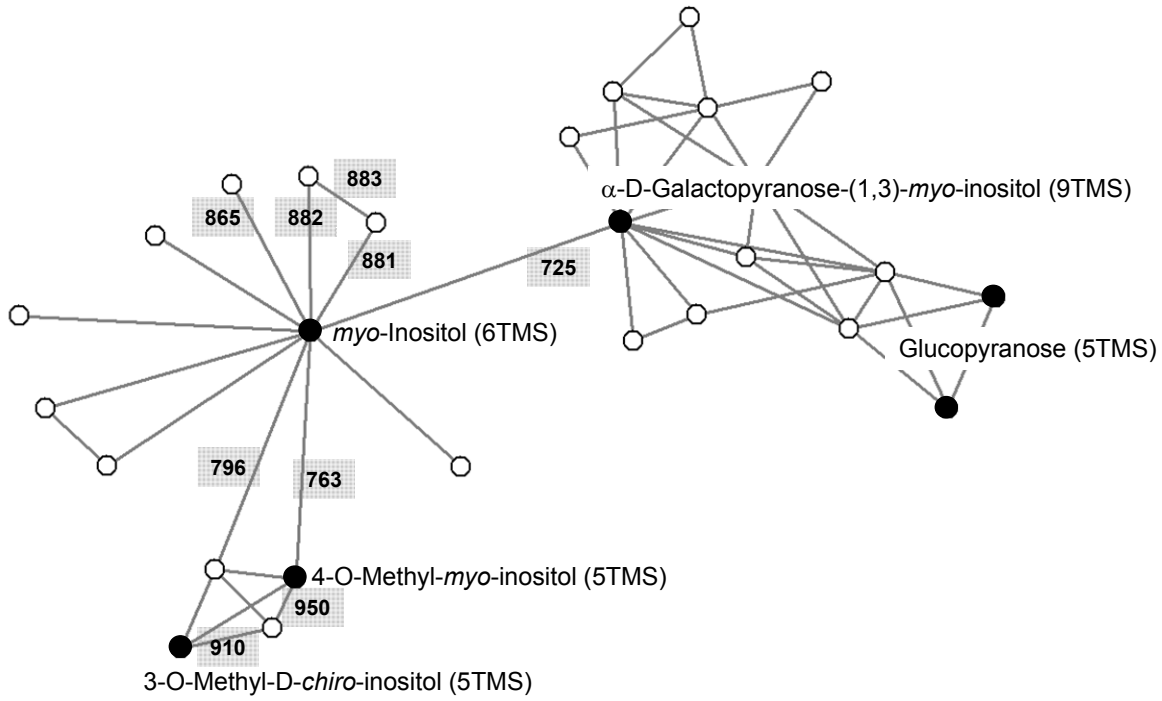


Figure 6 *Myo-Inositol centered proximity map of mass spectral similarity among identified and non-identified mass spectral tags (MSTs) from GC-MS profiles of complex biological sources and authenticated reference substances.* The search started at *myo*-inositol (6TMS) and used the symmetrical NIST08 matching factor to build the network of similar mass spectra. Open circles represent hitherto non-classified MSTs and connecting edges represent best mass spectral matches which are partially indicated by shaded boxes. Note that a path along the best pair-wise similarities groups links chemically related compounds and sorts non-identified MSTs into the scaffold of known compounds (Figure 3 of **Erban et al. 2007 [4]**).

Bottom-up Identification. With top-down identification approaching saturation, because the number of available pure reference compounds starts to become limiting, the bottom-up or *de novo* identification of MSTs will become the future focus of the metabolomics field. Even though novel developments of NMR technology, namely, dynamic nuclear polarization coupled to NMR analysis, promise enhanced instrument sensitivity by a factor >10,000 (Ardenkjaer-Larsen et al. 2003), the most direct strategy of compound identification from GC-MS profiles, namely linking preparative GC to structure elucidation

by NMR, may not become feasible for the time to come. Therefore, cheminformatic strategies which build on the increasing compendium of identified analytes within GMD have been pursued in this thesis. Hit lists as provided by the NIST08 search and comparison algorithm have been tested for compound classification. Such mass spectral hit lists allowed discovery of close chemical isomers within libraries of automatically deconvoluted MSTs from diverse biological matrices (**Wagner et al. 2003 [8]**). Automated deconvolution is mandatory for the processing of the large numbers of MSTs generated by GC-MS based profiling. While multiple deconvolution algorithms exist, the ChromaTof deconvolution tuned to GC-TOF-MS instruments (LECO, St. Joseph, MI, USA) currently appears to represent the best choice (Lu et al. 2008). Simple hierarchical clustering of such mass spectra from NIST hit lists revealed epimeric, geometric and positional isomers (**Wagner et al. 2003 [8]**). This potential has been demonstrated using the metabolite classes of polyhydroxyhexanoic acids, such as gulonic, gluconic and galactonic acid, and the family of caffeoylquinic acids, with its most nutritionally important member, chlorogenic acid. Navigation through proximity maps, based on the symmetrical matching factors, for example to dot product, provided by NIST allowed visualization and grouping of chemically related compounds, such as inositols, methylinositols, and inositol conjugates, for example galactinol (Fig. 6 and **Erban et al. 2007 [4]**). Finally, application of decision tree algorithms allowed integration of RI and MS data for compound classification and were successful in a feasibility study differentiating amino acids and sugars from the remaining library compendium (Hummel J, Strehmel N and Kopka J, personal communication). This study demonstrated that compound-class specific, generic mass fragments were utilized as relevant distinctive features for classification (Fig. 7). For example, the first decision in this pilot study was based on the relative abundance of m/z 89, equivalent to fragment $C_3H_9OSi^+$, typical of siloxy-moieties. Also, m/z 147 representing the structure $C_5H_{15}OSi_2^+$ was used by the decision tree algorithm. This fragment is indicative of vicinal siloxy-groups. Both, m/z 89 and m/z 147 are in agreement with the high occurrence of hydroxyl-moieties in sugars which are converted to siloxy-moieties by the inherent chemical derivatization of GC-MS based profiling. In addition, m/z 361, namely, $C_{15}H_{33}O_4Si_3^+$, is typical of di- and trisaccharides or hexose-conjugates as its origin are glucosidic pyranose-rings. Typical mass fragments of amino acids were also used. In detail, m/z 218, $C_8H_{20}NO_2Si_2^+$, containing C_1 and C_2 of alpha-amino acids and m/z 100 equal to $C_4H_{10}NSi^+$, a typical fragment of amine moieties, occurred in this analysis. All of these mass spectral interpretations were supported by respective mass shift

analysis of compounds prepared after *in vivo*-^{13}C-labelling of plants (cf. supplementary file 1 of **Huege et al. 2007 [7]** and paragraphs below).

In conclusion, the top-down approach of analyte identification within GC-MS profiles of complex samples has been implemented from pathway analysis (**Luedemann et al. 2004 [9]**) to internationally coordinated library establishment (**Schauer et al. 2005 [10]**) and web-browser based data dissemination (**Kopka et al. 2005 [11]**). These efforts contributed to and were in agreement with the metabolomics standards initiative (Sumner et al. 2007) and preceding standardization efforts (Bino et al. 2004). In the future this project will be continued and novel sources of authenticated reference substances will be accessed as these may be made available by commercial suppliers or by academic consortia. The bottom-up approach of analyte identification has initially been focused on application and testing of cheminformatic tools. Simple analyses such as hierarchical clustering (**Wagner et al. 2003 [8]**) and proximity maps (**Erban et al. 2007 [4]**) yielded already novel insights. Enhanced informatics tools will integrate automated machine learning structural elucidation (Fig. 7). The utilization and precision of decision tree algorithms is under investigation and will further be enhanced by integrating not only information on retention indices and mass spectral fragmentation, but also by addition of features that can be extracted from known metabolite and analyte structures. This structural information and respective files are currently integrated into the GMD database (Strehmel N, Hummel J and Kopka J, personal communication). In addition, attempts to estimate exact masses of fragments or molecular ions for sum formula prediction, for example the early study of Fiehn and co-authors (2000b), and *in vivo*-stable isotope labeling of organisms have already become useful (**Birkemeyer et al. 2005 [6]**, Hegemann et al. 2007) and will in the future allow integrated approaches of GC-MS based structure elucidation.

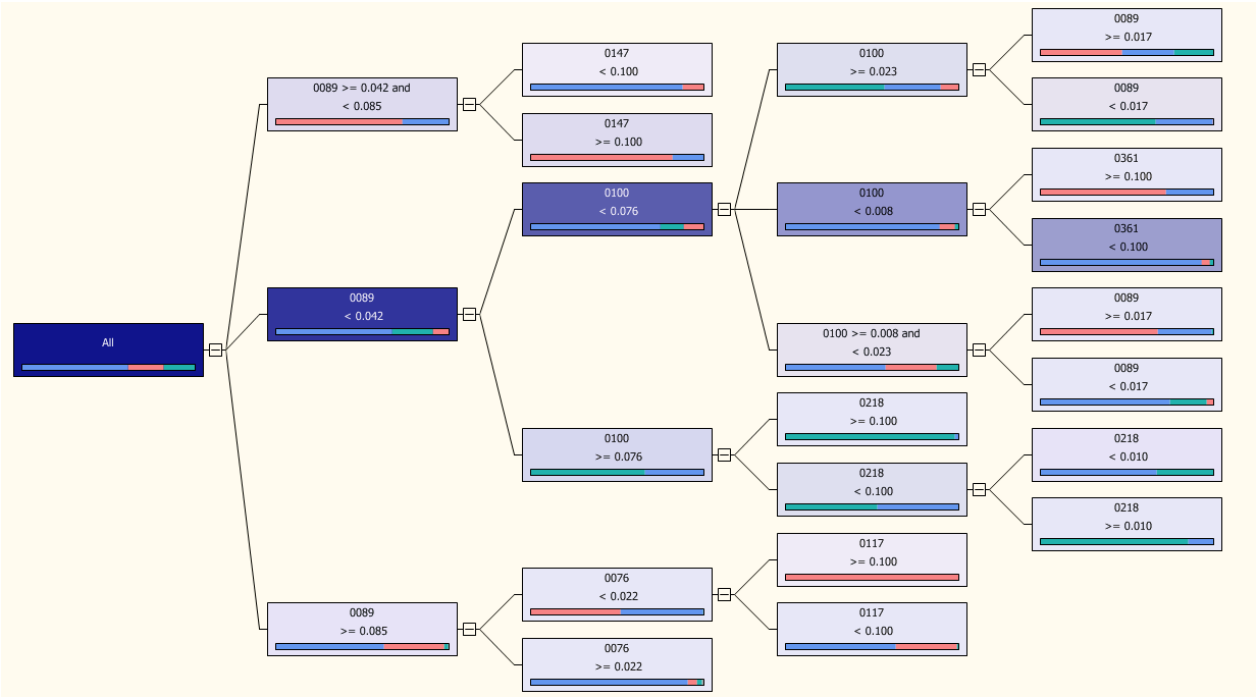


Figure 7 *Decision tree depicting automatically generated rules based on mass spectral properties which distinguish amino acid and sugar spectra from the remaining set of analytes in the Golm Metabolome Database (GMD).* Boxes show m/z (top), intensity threshold (middle) of the rule and the fraction of analyte classes (bottom) within each leaflet, namely, amino acids (green), sugars (red) and other analytes (blue). The background color intensity of each box is proportional to the total number of analytes within each leaflet. Note that the expected generic mass fragments of amino acids (m/z 100 and m/z 218) and of sugars (m/z 89, m/z 147, m/z 361) were automatically selected by the decision tree algorithm. In this feasibility study 61 % of the amino acids, 74 % of the sugars and 83 % of other compounds were correctly classified using cross validation (Hummel J, Strehmel N, Walther D and Kopka J, unpublished).

Automated Data Processing of Complex GC-MS Based Metabolite Profiles

Standardized and automated chromatography data processing was the initial bottleneck of high-throughput GC-MS based metabolite profiling. Two key functions are required, (i) the generation of a numerical data matrix aligned by mass and chromatographic retention which should provide a regular format ready for statistical analysis of several hundred GC-MS chromatograms and (ii) a repeatable and validated procedure for the recognition of analytes within complex biological extracts using custom reference libraries tuned to the need of this process. With this motivation a long-term software development, named TagFinder

(**Luedemann et al. 2008 [13]**), was initiated in the year 2004. The central aim was to provide automated tools which nevertheless allow user intervention and provide data suited for the purpose of uploading to statistics software or to a database of GC-MS profiles.

Contrary to the efforts of unifying the standards for data retrieval, data mining and interpretation, highly diverse and in part specialized software solutions for the GC-MS data pre-processing have been developed and published, because vendor software for GC-MS data acquisition and pre-processing traditionally focuses on targeted metabolite quantification. Such software solutions are not suited for non-targeted metabolomic analyses, because high-throughput and the discovery of novel metabolic components are not supported. Thus, automated mass spectral deconvolution has become a topic intensely explored in both, academic and commercial software development. Deconvolution performs comprehensive and non-biased extraction of mass spectra from GC-MS data files and provides relevant mass spectra for compound identification. The conventional approach of mass spectral deconvolution is based on information present within single chromatograms. For example, an early approach developed a so-called back-folding procedure for the mathematical enhancement of GC-MS based chromatographic curve resolution (Pool et al. 1996, Pool et al. 1997a, Pool et al. 1997b). In contrast, multivariate curve resolution (MCR) and its successor the hierarchical multivariate curve resolution (HDA) may represent the most advanced pre-processing tools. Information of multiple aligned chromatograms is integrated for deconvolution (Jonsson et al. 2004, Jonsson et al. 2005, Jonsson et al. 2006). However, MCR and also HDA are highly sensitive to the selection and number of analyzed chromatogram files and may fail using > 20 chromatogram files. In addition the targeted retrieval of selected fragment ions and the extraction of mass isotopomer distributions for flux analysis were not supported. The perhaps most widely spread tool is the Automated Mass Spectral Deconvolution and Identification System (AMDIS, http://chemdata.nist.gov/mass-spc/amdis/overview.html, Halket et al. 1999, Stein 1999), which is coupled to the standard mass spectral search and comparison software NIST08 (National Institute of Standards and Technology, Gaithersburg, MD, USA, http://www.nist.gov/srd/mslist.htm). AMDIS was initially designed for purely qualitative analyses. Quantitative information extracted by AMDIS is still poorly defined and hard to access. Finally, the commercial ChromaTof software (LECO, St. Joseph, MI, USA, http://www.leco.org/), combines deconvolution based identification and quantification of target metabolites. But this software is exclusive for the GC-TOF-MS systems of the vendor. While the ChromaTof deconvolution appears to be highly successful for compound discovery (e.g. **Wagner at al. 2003 [8]**, Lu et al. 2008), it

does not support the generation of an aligned data matrix. The data pre-processing comes at the price of software errors, such as partially deconvoluted MSTs, mixed or in other words chimeric MSTs, occurrence of artificial MSTs due to electronic noise, and erroneous MST duplications. Also quantitative errors of extracted fragment intensities are reported (e.g. **Lisec et al. 2006 [3]**).

Alternative to elaborate deconvolution algorithms simple peak retrieval and numerical matrix generation was pursued by software projects in academia. These tools were typically built on the comprehensive extraction of mass selective peak apex intensities. This approach is computationally less demanding and required fewer parameter settings compared to the, traditionally preferred, extraction of peak areas. A typical example of such software tools is the MetAlign collection of algorithms (http://www.pri.wur.nl/UK/products/MetAlign/). These also support mass alignment suitable for both, low mass-resolution GC-MS (e.g. Tikunov et al. 2005) and high mass-accuracy instruments (Bino et al. 2005, Vorst et al. 2005, America et al. 2006, Keurentjes et al. 2006, De Vos et al. 2007). Further software tools provide diverse options of numerical matrix generation and support non-targeted as well as metabolite targeted GC-MS analyses, for example XCMS (Smith et al. 2006), MeMo (Spasić et al. 2006), MathDAMP (Baran et al. 2006), MetaQuant (Bunk et al. 2006), the MSFACTs (Duran et al. 2003; http://www.noble.org/Plant-Bio/MS/MSFACTs/MSFACTs.html) and the respective refinement, MET-IDEA (http://www.noble.org/Plantbio/MS/MET-IDEA/index.html), or the progressive peak clustering approach of De Souza and co-authors (2006). First attempts have also been made at compound-targeted processing of the novel highly complex GCxGC-TOF-MS files (e.g. Sinha et al. 2004a). One unique tool, the BinBase (Fiehn et al. 2005; http://fiehnlab.ucdavis.edu/projects/binbase_setupx/), which combines GC-TOF-MS data analysis with a database application, collects and archives extracted MSTs and profiles. This tool, however, appears to be not publicly accessible.

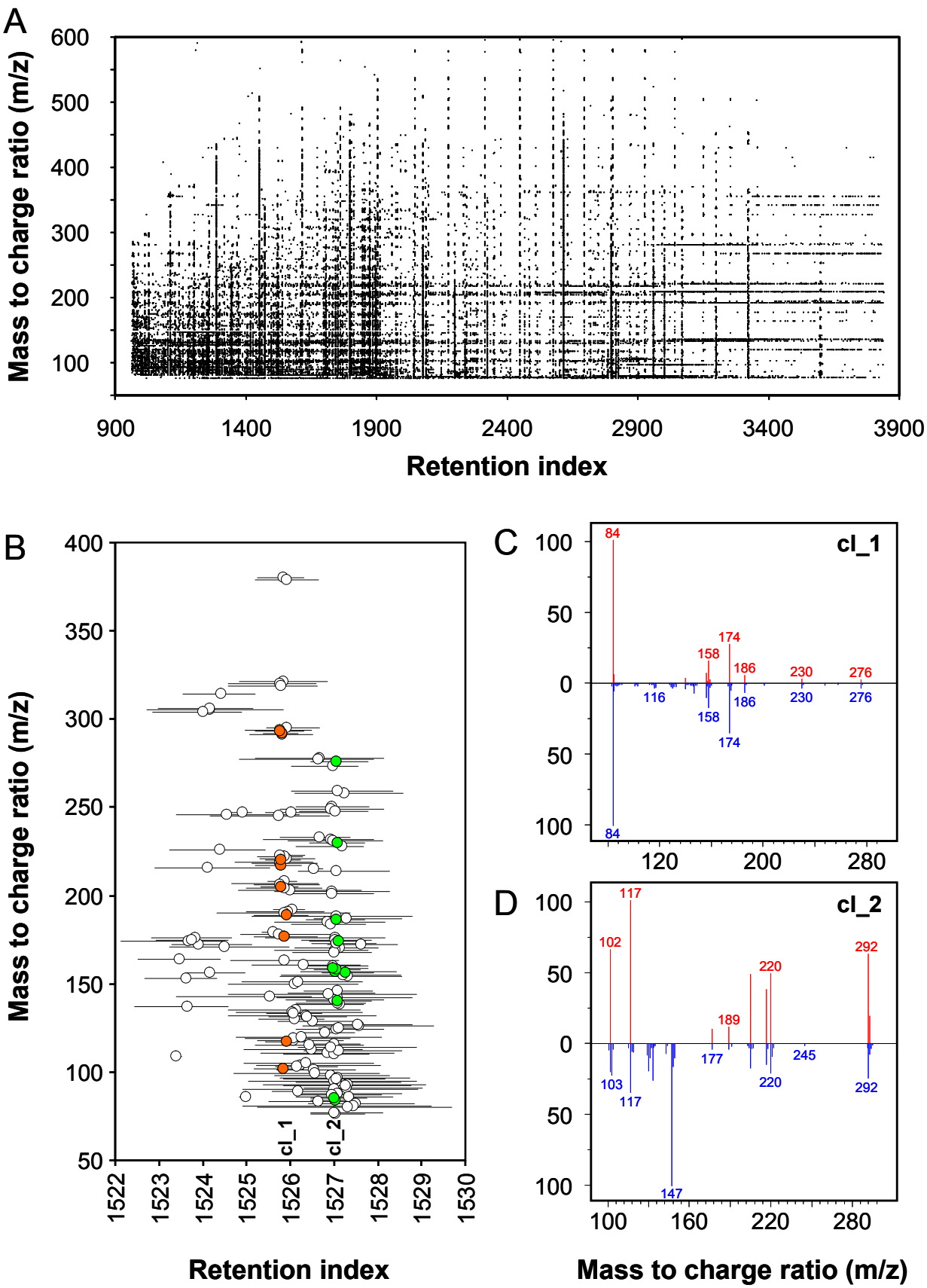

A
Mass to charge ratio (m/z)
600
500
400
300
200
100
900
1400
1900
2400
2900
3400
3900
Retention index
B
Mass to charge ratio (m/z)
400
350
300
250
200
150
100
50
cl_1
cl_2
1522
1523
1524
1525
1526
1527
1528
1529
1530
Retention index
C
cl_1
84
174
158
186
230
276
116
D
cl_2
117
102
292
220
189
177
245
103
147
100
160
220
280
Mass to charge ratio (m/z)

Figure 8 (<) *Visualization of the TagFinder-supported GC-TOF-MS data pre-processing procedure.* An experiment of 95 *Arabidopsis thaliana* leaf samples was processed. About 23,900 mass features, so-called mass tags, were found with frequency of occurrence set to >5% and detector response limited to >50 arbitrary units **(A)**. Each of these mass features is characterized by a retention index (RI) window with minimum, maximum and mean RI (circle and whiskers) observed across the complete set of 95 chromatograms **(B)**. These mass tags are grouped by TagFinder into so-called time groups by checking for overlapping RI windows. A zoom-in on a single time group **(B)** demonstrates the occurrence of multiple co-eluting mass tags observed in GC-MS experiments. A cluster analysis among the mass tags of a single time group selects those mass tags which are highly correlated according to Kendall`s or Pearson`s correlation applied to the observed intensities of all processed chromatograms. The resulting clusters, here named cl_1 (orange circles) and cl_2 (green circles), allow reconstruction of partial mass spectra which can then be matched to reference spectra by a target finding function of the TagFinder software. Here, cl_1 represents glutamic acid (2 TMS), with 11 matching mass features, RI deviation -0.07 % and an excellent match factor 955. Cluster cl_2 represents erythronic acid (4 TMS) with 9 matching mass features, RI deviation -0.30 % and an acceptable match factor 769 (Luedemann, A, Erban A and Kopka J, unpublished).

Even though a rich and diversified resource of free and commercial tools exists, none supports both, the standardized comprehensive numerical matrix generation including full mass isotopomer information and the standardized recognition of analytes. TagFinder software (**Luedemann et al. 2008 [13]**) solves this task and utilizes a generalized workflow comprising eight steps, (i) peak intensity retrieval and data import, (ii) annotation of sample identifier and sample information, (iii) alignment by retention index calculation and by full mass unit resolution, (iv) mass tag generation, (v) time group generation, (vi) time group clustering (Fig. 8), (vii) numerical data matrix generation, (viii) analyte recognition and selective mass tag assignment (**Luedemann et al. 2008 [13]**). The data matrix of each experiment is exported in tabular text format for uploading into statistical software tools, such as the TM4 open-source system (Saeed et al. 2003, Saeed et al. 2006), and supports a XML format for database upload. Compound recognition is performed separately for each experimental data matrix by matching reconstructed mass spectra and retention indices of time groups and clusters to reference library information. Several matching criteria are implemented, such as calculation of RI deviation, number of matched mass fragments, a numerical distance measure comparing vectors of normalized fragment intensities, similar to the NIST matching algorithm, and a novel distance measure comparing vectors of pair-wise fragment ratios (Luedemann A, Erban A and Kopka J, personal communication). These criteria are suitable to validate the quality of compound matching and are also exported either in tabular text or XML format.

In conclusion, this thesis has created a highly versatile software tool for the alignment of large GC-MS based metabolite profiling experiments into statistically accessible data matrices and has tested this tool in long-term routine applications. TagFinder has been applied to metabolic fingerprinting and profiling (e.g. Kopka et al. 2004, Kopka 2006a, Kopka 2006b), mass isotopomer ratio analysis (e.g. **Birkemeyer et al. 2005 [6]**) and to flux analysis (e.g. **Huege et al. 2007 [7]**). The matrix generation is directed by co-analysis of RI marker substances within each chromatogram. The simultaneous in-parallel analysis of chemically defined mixtures of reference compounds within each experiment is recommended for improved validation and analyte recognition (**Strehmel et al. 2008 [14]**). Automated extraction of quantitative data using pre-defined mass fragments, so-called time groups of mass fragments or clusters is implemented and has been linked to an analyte matching procedure. Both, the quantitative and the qualitative data types are database ready and a TagFinder extension for the pre-processing of GCxGC-TOF-MS data files is under development. Mass spectral matching within preset RI windows is provided and coupled to our custom reference libraries (**Kopka et al. 2005 [11]**, **Schauer et al. 2005 [10]**). The current and future efforts will explore the use of synthetically defined reference mixtures as well as matching thresholds (e.g. **Strehmel et al. 2008 [14]**) to establish rules for enhanced and automated compound recognition with minimum requirement of user interventions.

The Golm Metabolome Database (GMD)

Caused by the emergence of high-throughput technologies, databases and respective web-services have become essential tools for information storage and exchange in biological sciences as may be best exemplified by protein and nucleic acid sequence repositories. Databases are also ideally suited for empirical technologies which generate invariant data formats, such as large scale transcriptomic or metabolomic screenings. These approaches create qualitative or quantitative data sets that gain additional value when results are compared between experimental conditions and laboratories. Today, biological databases are ubiquitous and growing in numbers. The molecular database collection of the year 2005 lists 719 biology related databases which are freely available to the public and reports an annual increase of 171 during 2005 (Galperin 2005). Meta-databases start to facilitate guidance and access to such computer-readable data (Cary et al. 2005, Bader et al. 2006). For example, 190 web-accessible resources of biological pathways and networks were available in 2006 (Bader et al. 2006). The metabolomics community does not yet support a dedicated central data

repository and the metabolome scientist has to aggregate and integrate relevant data from a high number of specialized commercial and non-commercial databases. These useful metabolomic resources were recently been reviewed by Arita (2004) and Mehrotra and Mendes (2006). The relevant information covers pathway knowledge, chemical substance information, related publications, and cheminformatic information specific to the applied profiling technologies, e.g. NMR or mass spectrometry. Given the highly diverse analytical technologies the development of application focused databases can be envisioned (e.g. Moco et al. 2006), which will communicate thorough common information resources, such as genome based pathway and metabolome reconstructions, and will require common interchange data formats, for example representations of metabolite profiles by fold-changes of pool size compared to defined biological and chemical reference material. The field as a whole, however, may still be considered in its early stages and standardized data formats may arise in the future from efforts, such as the first Metabolomics Standards Workshop (August 2005, Bethesda, MD, USA).

The role of GMD. GMD started as a collection of annotated and non-annotated but repeatedly observed mass spectra from defined biological samples which was extended to include information on chromatographic retention behavior for enhanced manual analyte recognition (**Wagner et al. 2003 [8]**). The subsequently developed concept of mass spectral tags, MSTs (cf. **Desbrosses et al. 2005a [20]**, Kopka 2006a, Kopka 2006b) allowed the handling and referencing of yet non-identified metabolic components from GC-MS profiling experiments and will facilitate future identification, once novel pure and authenticated reference substances may become available. In the following many expert laboratories in the field complemented and contributed to the initial collection (**Schauer et al. 2005 [10]**). The first phase of GMD development culminated in a web-interfaced database for public queries and downloads of the library content (**Kopka et al. 2005 [11]**). The initial role of GMD was the dissemination of reference mass spectra and information on biologically relevant GC-MS analytes. Thus, GMD complemented the large NIST and Wiley spectral libraries which include only a limited fraction of the compounds frequently observed in GC-MS profiling experiments. GMD has served the metabolomics community since 2005 by providing a common basis for metabolite recognition and has since received 143 citations, 09/2008 (**Schauer et al. 2005 [10]**, **Kopka et al. 2005 [11]**).

Because of the wide acceptance GMD is currently enhanced in cooperation with expert bioinformaticians to a server-based relational database (Hummel et al. 2008). The central database object of GMD is the MST, namely, mass spectra comprising elements of mass to

charge ratio, fragment abundance and chromatographic retention linked to GC-MS method descriptions. MSTs link the workflow of metabolite identification to the high-throughput metabolite fingerprinting and profiling workflow. Both workflows are represented within the design of a relational database model and are described in detail by Hummel and co-authors (2008). In the following a short description of the GMD database model is given.

The metabolite identification workflow. The metabolomic identification process links analytical readings, here MSTs generated by GC-MS instruments, to the compound structure of metabolites. GMD models the top-down identification process which starts with a metabolite structure. For this purpose GMD contains a metabolite object which is linked through external database identifiers to the most frequently used metabolite or compound repositories, namely, KEGG (http://www.genome.jp/kegg/) with more than 14,000 metabolites, PubChem (http://pubchem.ncbi.nlm.nih.gov/) providing ~8 million substances (Kremsky 2005, Shang and Tan 2005) and the commercial Scifinder database (http://www.cas.org/SCIFINDER/) of the Chemical Abstracts Service (CAS) comprising more than 29 million synthetic and natural chemicals, e.g. (Schwall and Zielenbach 2000, Whitley 2002, Ben Wagner 2006). GMD links the metabolite object to an object representing commercially available and purchased reference substances. These reference substances generate the chemically derivatized analytes. Finally, the respective analyte object is linked to the MSTs which were observed by GC-MS analysis. The objects metabolite, reference substance and analyte have the common properties of chemical substances and detailed structure information is embedded as a conventional mol-file or as a more recently developed InChI code (IUPAC, International Chemical Identifier; http://www.iupac.org/inchi/). Structure annotations allow prediction of expected analytes and of MST properties, for example molecular ions and the preferred electron impact induced fragmentations of GC-MS analyses. Thus, GMD documents the complete top-down metabolite to MST identification process.

The metabolite fingerprinting and profiling workflow. This workflow is supported by the chromatography data processing and XML export files from the TagFinder software. Respective files, representing the GC-MS data matrix and sample information of each experiment are uploaded into GMD. TagFinder software then guides the user through a matrix specific MST and analyte recognition process. The criteria suitable to validate the quality of the compound recognition, such as spectral matching factors and RI deviation, are uploaded into GMD and used to link profiles to MSTs and, thus, to metabolites. The GMD infrastructure is currently tested using the large number of potato tuber profiles generated by

the BMBF QuantPro program, "Innovative diagnostic tools to optimize potato breeding: systemic analysis of cellular processes and their relation to plant internal oxygen concentrations" (Hummel J, Steinfath M, Strehmel N and Kopka J, personal communication).

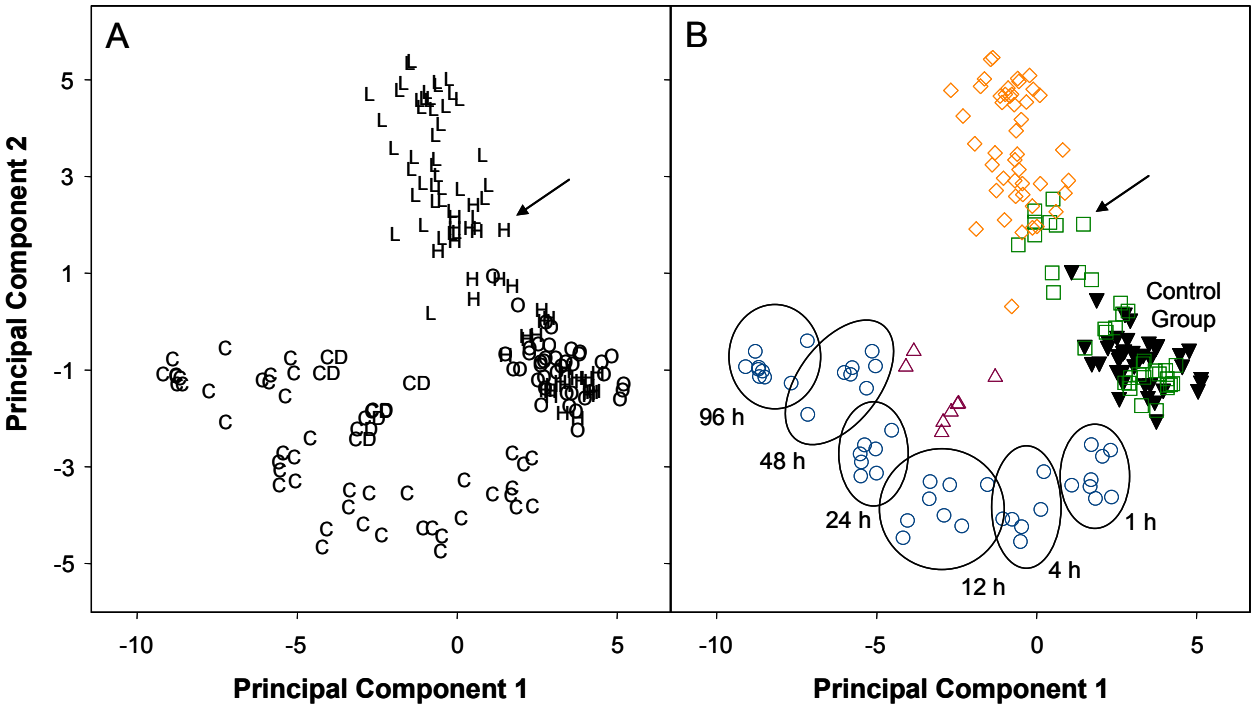
Table 1 The Golm Metabolome Database (GMD) currently harbors retention indices and mass spectra of 2,437 non-redundant MSTs from 1,632 commercially available reference substances. The transfer of retention index (RI) properties between chromatography variants of the contributing laboratories is demonstrated in this analysis (cf. Table 3 of **Strehmel et al. 2008 [14]**)[a]

	Variant 1	Variant 2	Variant 3	Variant 4	Variant 5	Variant 6	Variant 7	Variant 8
Column Brand	VF5	RTX5	RTX5	DB5	VF5	DB5	RTX5	Eq5
Temperature Ramp (°C/min)	9	9	15	40	6	6	5	3
Gas chromatograph	6890N	6890N	6890N	6890N	Trace GC ultra	Trace GC	GC 8000	Trace GC
Aquisition rate (scans/s)	20	20	6	30	2	2	1.67	2
Mode of Detection	TOF	TOF	TOF	TOF	Q	Q	Q	TRAP
A	Number of Paired Analytes							
Variant 1	488	348	274	179	157	175	318	65
Variant 2		931	623	209	226	244	518	96
Variant 3			964	184	206	192	437	93
Variant 4				299	151	127	197	71
Variant 5					264	154	224	77
Variant 6						324	190	82
Variant 7							961	96
Variant 8								103
B	Correlation Coefficient (r^2)							
Variant 1	-	0.99984	0.99971	0.99945	0.99914	0.99951	0.99926	0.99956
Variant 2	0.99983	-	0.99979	0.99951	0.99910	0.99961	0.99932	0.99962
Variant 3	0.99971	0.99979	-	0.99950	0.99973	0.99977	0.99977	0.99945
Variant 4	0.99938	0.99945	0.99943	-	0.99880	0.99983	0.99886	0.99976
Variant 5	0.99916	0.99913	0.99974	0.99904	-	0.99987	0.99974	0.99962
Variant 6	0.99951	0.99961	0.99977	0.99983	0.99987	-	0.99981	0.99975
Variant 7	0.99925	0.99931	0.99977	0.99910	0.99974	0.99981	-	0.99978
Variant 8	0.99956	0.99962	0.99945	0.99976	0.99963	0.99975	0.99978	-
C	Standard Deviation ($RI_{predicted}$ - $RI_{determined}$)							
Variant 1	-	7.42	9.21	12.71	16.05	9.54	16.04	9.55
Variant 2	7.50	-	8.00	12.01	17.13	9.33	15.00	8.20
Variant 3	9.30	8.10	-	12.12	9.18	7.83	9.29	9.58
Variant 4	13.64	12.89	13.08	-	19.81	5.69	19.58	7.36
Variant 5	15.93	16.82	9.14	17.56	-	5.11	9.36	9.22
Variant 6	9.56	9.37	7.73	5.66	5.11	-	6.75	6.83
Variant 7	16.19	15.04	9.27	17.17	9.28	6.70	-	6.80
Variant 8	9.07	7.82	9.10	6.98	8.72	6.50	6.53	-
D	Standard Deviation ($RI_{predicted}$ - $RI_{determined}$ [% of $RI_{determined}$])							
Variant 1	-	0.38	0.42	0.52	0.59	0.55	0.64	0.52
Variant 2	0.39	-	0.42	0.47	0.61	0.64	0.60	0.44
Variant 3	0.43	0.44	-	0.54	0.46	0.51	0.43	0.57
Variant 4	0.53	0.48	0.54	-	0.66	0.35	0.69	0.40
Variant 5	0.59	0.62	0.45	0.62	-	0.29	0.40	0.46
Variant 6	0.56	0.67	0.50	0.35	0.29	-	0.37	0.38
Variant 7	0.64	0.60	0.42	0.64	0.40	0.36	-	0.36
Variant 8	0.50	0.43	0.55	0.38	0.44	0.37	0.35	-

[a] All included method variants were based on 5%-phenyl-95%-dimethylpolysiloxane or equivalent stationary phases and were operated at 1 mL/min constant helium flow. Column brand, temperature programming and mass spectral detection varied as indicated. (A) Number of paired analytes, which were used for 3rd order polynomial regression, (B) regression coefficients, r^2, (C) accuracy of prediction as characterized by standard deviation of residual errors, $RI_{predicted} - RI_{determined}$, and (D) accuracy of prediction as characterized by standard deviation of residual errors expressed as percent of $RI_{determined}$. Note that due to the algorithm resulting matrices B and C are not exactly symmetrical; horizontal variants were used to predict RIs of the variants listed vertically.

Figure 9 (>) *Principal component analysis covering 38.5% and 21.9 % of total variance in a data set of leaf metabolite profiles from Arabidopsis thaliana eco-type Columbia.* Plants were environmentally challenged (i) by high light (L, diamonds; long-term adaptation to 560 and 850 µE/m² compared to a control at 120-150 µE/m²), (ii) by high temperature (H, squares; up to 4 h at 40°C compared to a control at 20°C; **Kaplan et al. 2007 [18]**) and (iii) by low temperature (C, circles; up to 96h at 4°C compared to a control at 20°C, **Kaplan et al. 2004 [17]**). Different formatting highlights environmental challenge (A) and time course compared to the control group (B). Note that high light and high temperature response exhibit a partial overlap (arrows). Cold de-acclimated plants (CD, triangles; 24h reversion to 20°C after 96 h at 4°C) show the existence of a metabolic memory after reversion to optimum temperature conditions (cf. Figure 1 of Steinhauser and Kopka 2007). C, cold; CD, cold de-acclimation; H, heat; L, light; O, control samples representing optimal growth conditions.

In conclusion, GMD has been created as a GC-MS focused public database for the successful dissemination of reference mass spectra and retention information for reproducible analyte recognition in the metabolomic field (**Schauer et al. 2005 [10]**, **Kopka et al. 2005 [11]**). Subsequently GMD has been extended to a relational database (unpublished) which harbors a bipartite data model suited for both, the metabolite identification and the profiling workflow. The added value of GMD has so far been exploited with a focus on top-down identification (Fig. 4-7). First analyses of the database content have begun and will provide empirical thresholds for MST recognition (**Strehmel et al. 2008 [14]**). This study uses the redundant entries within the GMD collection for RI prediction between typical variants of the GC-MS based metabolite profiling method and estimated 0.5 -1.0 % RI precision for compound recognition (Tab. 1). Because these thresholds will not solve all matching ambiguities in complex samples, the co-analysis of reference substances with each GC–MS profiling experiment was recommended. The composition of such defined reference mixtures may best approximate or mimic the quantitative and qualitative composition of the biological matrix under investigation. Future efforts will continue metabolite identification, utilize GMD information to classify and characterize yet non-identified MSTs and most importantly focus on routine uploading of TagFinder processed profiling experiments. Thus, a compendium of metabolite profiles which describe mutant and natural metabolite phenotypes in response to nutritional and environmental stress conditions is envisioned for future comparative metabolic pattern analysis (Fig. 9).



Enhanced Metabolite Profiling using Mass Isotopomer Ratios (ITR)

GC-MS based metabolite profiling has high inherent reproducibility. The quantitative dynamic range typically covers three orders of magnitude or more (Fig. 10A). Technical reproducibility can be below 10 % relative standard deviation (RSD) unless measurements are near detection limits (Fig. 10B). Metabolite recoveries, mostly in the 70-130 % range, and linear range of quantification have been established in early studies, using exemplary biological sample types (e.g. **Fiehn et al. 2000a [1]**, **Roessner et al. 2000 [2]**). However, two sources of technical artifacts were uncovered: (i) Metabolites may generate two or more alternative analytes, for example different degree of trimethylsilylation or Z- and E-products of methoxyamination. The ratios of such analytes are mostly stable, but may vary if the composition of the investigated biological matrix affects reagent surplus or reagent reactivity. (ii) The chosen biological matrices may also be resilient to reproducible extraction or may have low and non-reproducible metabolite recoveries. The best solution to the standardization of such detrimental effects is the use of chemically synthesized, stable isotope labeled, quantitative internal standard substances. This procedure was suggested for routine standardization of exemplary compounds (**Fiehn et al. 2000a [1]**) and was subsequently found to substantially enhance analytical RSD (e.g. Gullberg et al. 2004). However, this gold

standard of relative quantification renders metabolite profiling experiments expensive and restricted only to those compounds which are accessible to chemical mass isotopomer synthesis.

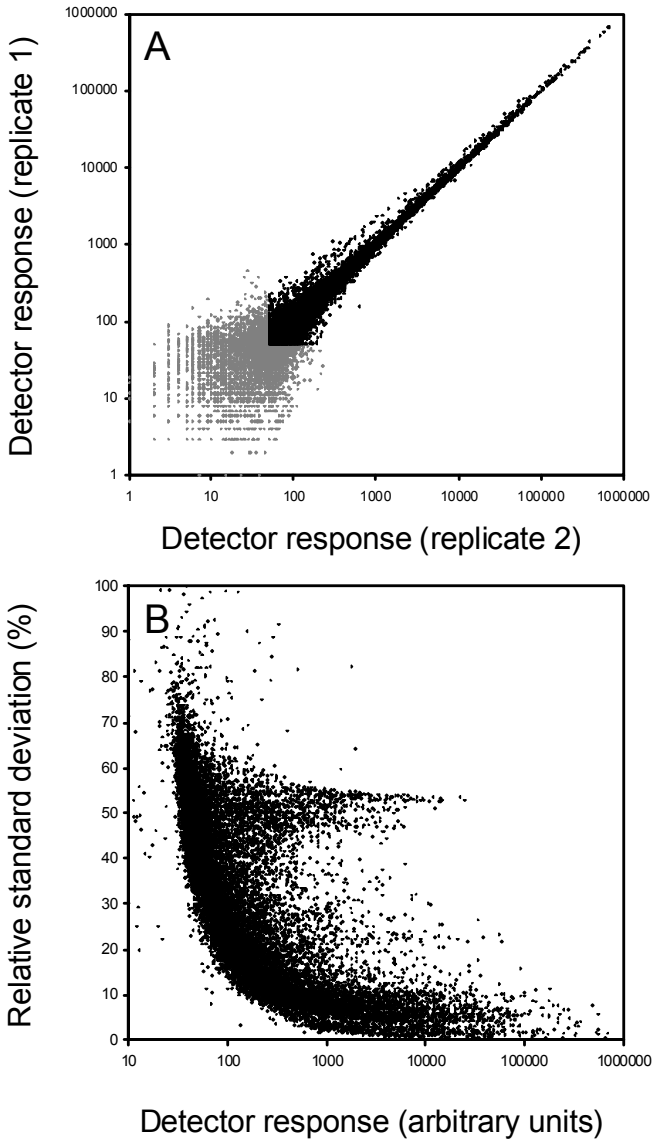


Figure 10 *Technical reproducibility of GC-TOF-MS based metabolite profiling following the method of Erban et al. (2007)* ***[4]***. Technical replicate analyses of *Oryza sativa* cv. Nipponbare leaf material was performed by preparation of 29 identical aliquots from a stock of deep-frozen, homogenized powder. The detector response of all common mass features from two replicate analyses **(A)** demonstrates the high degree of reproducibility. The relative standard deviation (RSD) of each mass feature across all replicate measurements **(B)** defines the optimum range of relative quantification. Note that intense features representing metabolites are typically below 10 % RSD and that the population of intense features at 50-60 % RSD is caused by reagent contaminations. GC-TOF-MS chromatograms were processed by TagFinder software. Gray format in subfigure A indicates the response cutoff suggested for the processing of this experiment (Erban A and Kopka J, unpublished).

A solution to the efficient standardization of metabolite profiling experiments was found by combining *in vivo*-stable isotope labeling of whole organisms with relative (**Birkemeyer et al. 2005 [6]**) or absolute (Mashego et al. 2004, Wu et al. 2005) quantification of metabolite pool sizes. First the yeast, *Saccharomyces cerevisiae* (**Birkemeyer et al. 2005 [6]**), then higher plants, such as *Arabidopsis thaliana* or *Oryza sativa* (**Huege et al. 2007 [7]**), were used as models for technology development. In pre-experiments ^{2}H, ^{15}N and ^{13}C labeling were compared. Labeling by $^{2}H_2O$ (D_2O) proved impossible as this isotopologue of water was toxic to higher plants (Erban A and Kopka J, unpublished). The use of high ^{15}N-enrichment proved feasible in higher plants, but - inherent to the chosen element - only the sub-fraction representing the N-metabolome was labeled (Engelsberger et al. 2006). Finally, the use of the ^{13}C-isotope exhibited broadest metabolite coverage coupled to high mass shifts. With U-^{13}C-glucose as exclusive carbon source yeast was readily ^{13}C-labeled. Only minor effects on metabolite pool sizes were detected and viability appeared unaffected (**Birkemeyer et al. 2005 [6]**). Subsequently, using a CO_2 controlled plant growth chamber built by GMS Gaswechsel-Messsysteme GmbH, Berlin, Germany (http://www.gms-biobox.de/Biobox_english/biobox_english.html), a method for the *in vivo*-labeling of higher plants in hydroponic cultivation was established (**Huege et al. 2007 [7]**). Plants were fully viable and 90 atom% or higher ^{13}C-enrichment was achieved within the soluble metabolite pools accessible to GC-MS profiling. Full stable isotope labeling of higher plants was also reported in parallel studies using alternative *in vivo*- or *in vitro*-labeling strategies (Dueck et al. 2007, Hegeman et al. 2007).

With efficient *in vivo*-labeling of yeast and higher plants in place, a method for the enhanced relative quantification by GC-TOF-MS profiling appeared to be feasible and was developed. This method uses identical aliquots of ^{13}C-labeled biological reference material, namely, whole cell preparations, for comprehensive, stable isotope based, internal standardization. The ratios of monoisotopic ^{12}C- and ^{13}C-mass signals representing identical chemical fragmentation products or molecular ions are used to estimate fold changes of pool sizes compared to the spiked labeled biological reference material. Both, identified and non-identified analytes can thus be tested for recovery. After patent submission (Luedemann et al. 2005) the approach was published as a tool for enhanced GC-TOF-MS based metabolite profiling (**Birkemeyer et al. 2005 [6]**) and then judged to be part of the next wave in metabolome analysis (Nielsen and Oliver 2005). When applied to the profiling of the intracellular metabolites of yeast the enhancement by stable isotope ratio profiles becomes evident (Fig. 11).

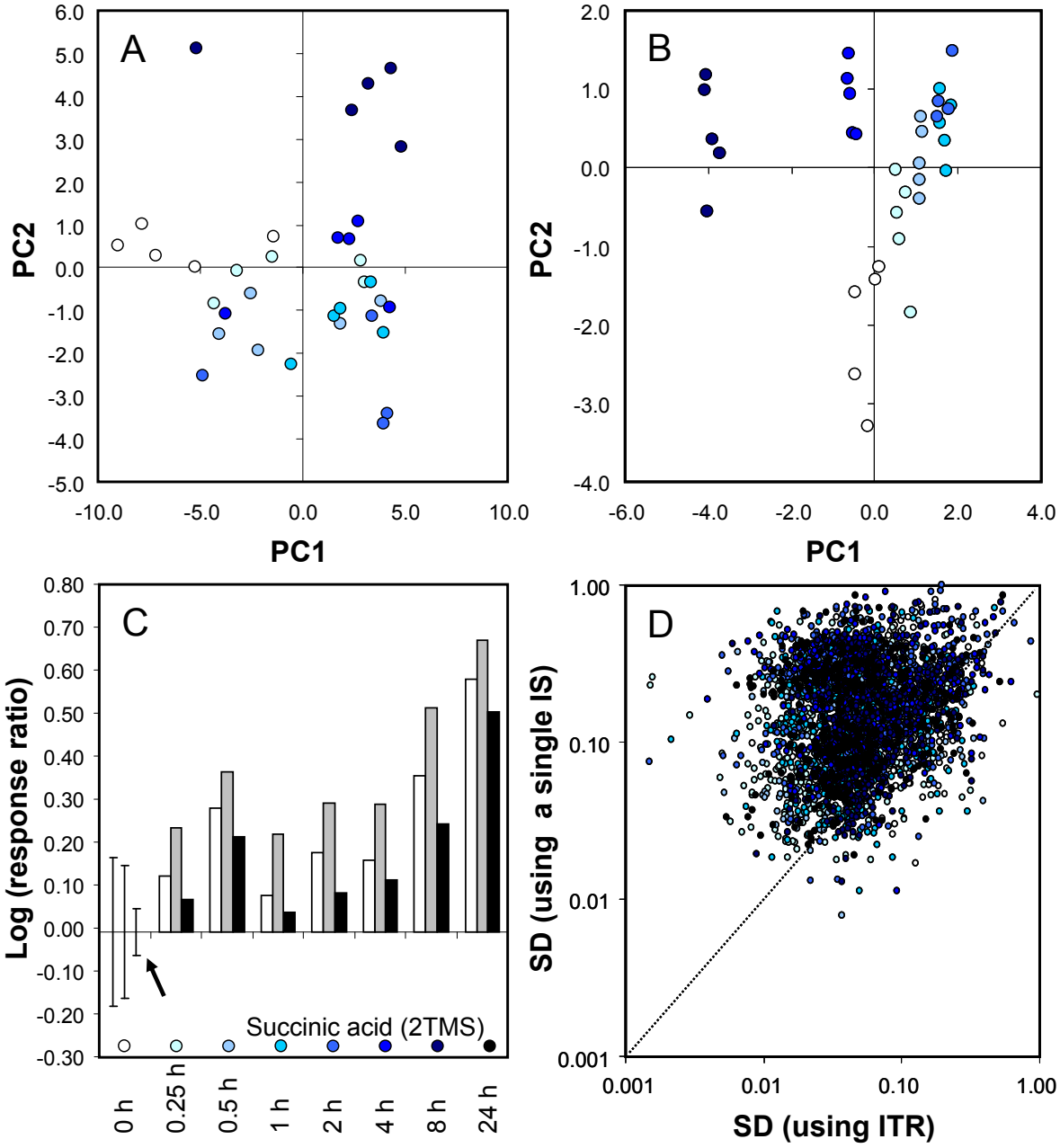


Figure 11 *Improved accuracy of ITR enhanced GC-TOF-MS based metabolite profiling.* A yeast, *Saccharomyces cerevisiae*, cell culture was grown to OD 1.0 and then subjected to a 10 °C cold stress. A time course was sampled into -60 °C cold methanol at 70 % final concentration using equal aliquots of pre-sampled ^{13}C-labeled yeast cells as internal standard for mass isotopomer ratio (ITR) metabolite profiling. Principal component analyses of the same metabolite set compare normalization by a single internal standard (IS), namely, ribitol **(A)**, to ITR enhanced normalization **(B)**. Note that that the overall technical variation is substantially reduced by ITR. The gain in precision is caused by reduced standard deviation **(C, arrow),** but does not alter substantially mean metabolite profiles. This finding is exemplified using a representative example; normalization by cell number (white bars), normalization using a single IS (gray bars) and normalization using ITR (black bars), standard deviation (n = 5) is indicated at t_0 **(C)**. A comprehensive comparative analysis **(D)** demonstrates that - with few exceptions - all ITR normalized mass features have equal or reduced standard deviation (SD). The time course is color coded **(cf. C)**. (Strassburg K and Kopka J, unpublished).

In conclusion, this thesis has established the *in vivo*-stable isotope labeling of higher plants and of yeast which was chosen as an initial model organism for technology development. This achievement has been employed to enhance the accuracy of GC-TOF-MS based metabolite profiling by quantifying mass isotopomer ratios, ITRs (Luedemann et al. 2005; **Birkemeyer et al. 2005 [6]**). ITR enhanced metabolite profiling is laborious compared to traditional metabolite profiling. For this reason, ITR enhanced profiling will remain restricted to resilient biological matrices but it may prove to be superior for future metabolic modeling approaches. The upcoming studies in my laboratory will focus on the utilization of ITR enhanced metabolite profiling for the analysis of trace compounds from plant material. Such efforts will build on the previous experience of inherent technological limitations of phytohormone profiling. The GC-MS based profiling of phytohormones was explored in my laboratory (**Birkemeyer et al. 2003 [5]**), but this method required several enrichment steps and highly accurate control of phytohormone recovery. Therefore, the technology was essentially limited by availability of stable isotope labeled internal standards (**Birkemeyer et al. 2003 [5]**). The required complex mixtures of chemically diverse phytohormones were excessively expensive or a matter of specific customized chemical synthesis. Now reference mixtures which should contain phytohormones can be provided as part of *in vivo*-labeled biological reference material.

Towards Combined Metabolite Pool Size and Flux Analysis

As the causal physiological interpretation of metabolite pool size changes is limited, mostly due the polygenic nature and multiple levels of metabolic control, flux analyses may be seen as the key technology for a more detailed functional insight into metabolic processes. GC-TOF-MS technology allows the measurement of both, pool sizes using mass fragment abundance and flux using the mass isotopomer distribution, to determine the fractional enrichment of stable isotope labeled elements over time. In short GC-TOF-MS metabolite profiling technology has the potential of combined pool size and flux analyses and in combination with ^{13}C-tracing experiments is suited to tackle the physiology and dynamics of carbon partitioning in photosynthetic organisms. The partitioning of carbon within the plant is perhaps the most essential topic in modern biology moving towards systems analysis. Insight into this process will substantially contribute to our functional understanding of plant acclimation responses to nutritional or environmental cues and the underlying gene functions.

The tracing of assimilated CO_2 is the most direct approach towards monitoring carbon partitioning in photosynthetic organisms. For this reason the labeling of plants with CO_2 with both, radioactive (e.g. Calvin M, 1956; Calvin M, 1961) or stable isotope tracing (e.g. Schaefer et al. 1975, Schaefer et al. 1980, MacLeod et al. 2001, Schwender et al. 2004) has been applied utilizing the main entry points of CO_2 into plant metabolism, namely, ribulose-1,5-diphosphate carboxylase (EC 4.1.1.39) and phosphoenolpyruvate carboxylase (EC 4.1.1.31). CO_2 tracing studies yielded ground-breaking biological insights into photosynthetic carbon assimilation and, thus, into the essential life-sustaining physiological mechanisms on earth.

As the steady state assumption may not hold true for most photosynthetic systems the focus of the most recently started technology development project in my laboratory was set on enabling dynamic flux estimations of soluble metabolite pools (e.g. Ratcliffe and Shachar-Hill, 2006). This approach is currently discussed for the phenotypic analysis of gene function in higher plants (e.g. Fernie et al. 2005, Baxter et al. 2007) or cyanobacteria (e.g. Shastri and Morgan 2007) and first lesson lessons have already been learned in my laboratory.

CO_2 partitioning into polar metabolite pools of higher plants. The combined estimation of pool size and flux represents perhaps one of the most demanding technology developments. Kinetic measurements of higher plants are best performed on homogenously grown populations of individuals so as to establish a high resolution time course and to address the plant to plant variation of identical genotypes. Also induction of wound reactions and positional or developmental effects are avoided which would obscure kinetic measurements if the same plant individual is successively sampled several times. Population sizes of 60 plants or 2 x 30 for pair-wise comparisons were found sufficient. In my laboratory the chase after 4-5 weeks ^{13}C-labelling was monitored because the chase can be easily performed by exposure to ambient CO_2 and air. This asset enables good plant pre-acclimation and rapid sampling with minimal environmental perturbation. Except for air humidity environmental factors were kept constant during the chase period and representative enzyme activities remained unchanged even 3 days after initiation of the chase period. As was expected, the experimental through-put was low.

A first simplified data processing scheme was established starting with TagFinder software processing of the complex mass isotopomer distribution data from GC-TOF-MS chromatograms. Assuming homogenous metabolic labeling by $^{13}CO_2$ feeding, isotopomer distributions were first converted to fractional enrichments (atom %) using the sum formulas of molecular ions and mass fragments, of electron impact induced fragmentation reactions

(**Huege et al. 2007 [7]**). The second assumption made was a first reaction order of the initial chase kinetics. Together these assumptions allowed calculation of ^{13}C-half life in each of the monitored metabolite pools. This concept may represent the basis of a potential measure for metabolic C-partitioning. Different organs and parts of the same plant, namely, root and shoot, were monitored in parallel. Experiments had to be limited to the initial reaction kinetics, because the chase kinetics turned increasingly variable under prolonged conditions (Fig. 12, insert).

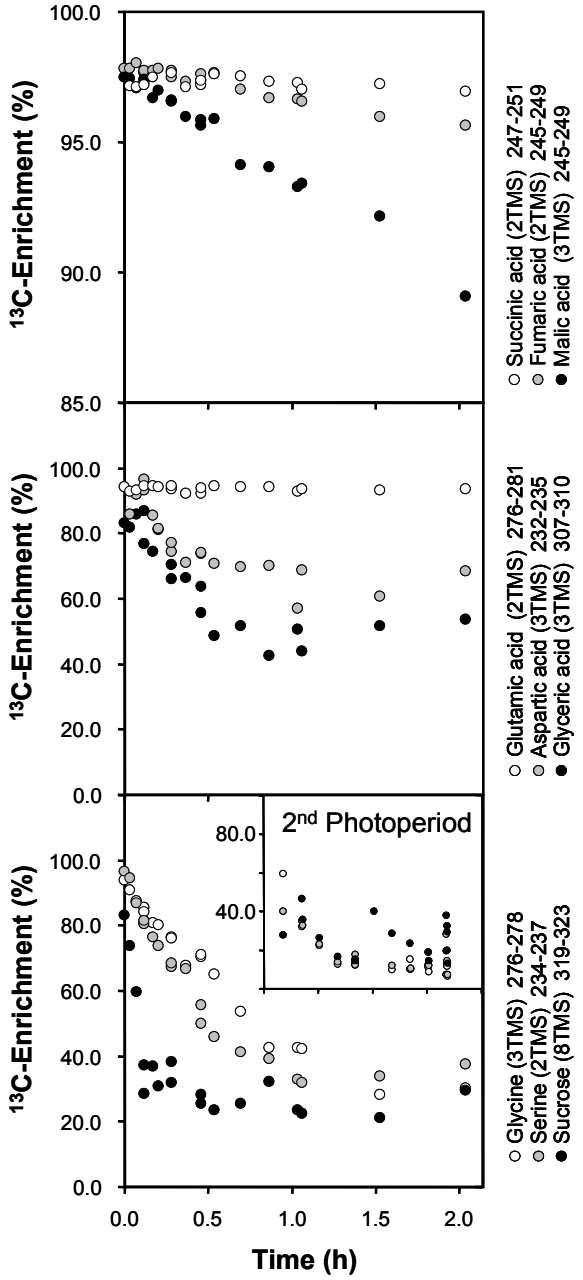


Figure 12 *The short-term kinetic behavior of ^{13}C-isotope dilution into exemplary metabolite pools of Arabidopsis thaliana Col-0 shoot upon exposure to ambient CO_2.* ^{13}C-decay was traced during the initial 2 h of the first light period and throughout the second photoperiod, insert (cf. Figure 7 of **Huege et al. 2007 [7]**).

Perhaps the most striking and unexpected observation was a biphasic chase response. After initial first reaction order kinetics a phase of "quasi steady states" at fractional enrichments elevated above the expected ambient composition was apparent in many metabolite pools. This finding may be indicative of a mixed use of newly assimilated carbon resources and turn-over of previously assimilated carbon stores (Fig. 12).

CO_2 partitioning analysis using cyanobacterial cultures. Flux analyses of higher plants, as described above, will always be subject to the influences of individual phenotypic variation of identical genotypes. Therefore, the analysis of cyanobacterial cultures was pursued in co-operation with an expert laboratory in cyanobacterial physiology (Eisenhut et al. 2008). This study led to the discovery of a partial metabolic phenocopy of on one side cyanobacterial cells shifted from high to low CO_2 supply compared to mutants with defined blocks in the photorespiratory metabolism including an unexpected increase of the C/N signaling molecule, 2-oxoglutarate. First dynamic flux studies were performed to understand this finding using photoautotropic cell cultures of the *Synechocystis* sp. PCC 6803 reference strain and photorespiratory mutants (data not shown). A ^{13}C-pulse was applied to cultures pre-grown under ambient conditions at defined cell density by adding a large excess of ^{13}C-labeled carbonate. The subsequent chase was set 60 min after pulse by substituting carbonate free medium after centrifugation and aeration with ambient air (Fig. 13). Cells were sampled by rapid, 20-30 sec, harvest onto a filter membrane and immediate shock freezing into liquid nitrogen.

The initial results using high ^{13}C-carbonate medium demonstrated fast and immediate ^{13}C-fractional enrichment of photoautotrophic *Synechocystis* sp. PCC 6803 cultures occurring as expected (e.g. Shastri and Morgan 2007) at 10-15 min with maximum enrichment reaching 92 atom % and 96 atom % in 3-phosphoglycerate and phosphoenolpyruvate pools, respectively (Fig. 13). Surprisingly, in all experimental repetitions the 3-phosphoglycerate pool exhibited a slightly lower apparent fractional enrichment than phosphoenolpyruvate. This observation may indicate and artifact caused by fast assimilation of ambient CO_2 during filtration harvest of the cell cultures. This potential artifact is currently under investigation so as to rule out or substantiate a physiological cause of this finding.

All observed responses of our pilot study were clearly non-steady state and in agreement with the dynamic flux concept, as the experimental intervention necessary for the carbonate pulse chosen for this experiment was designed to suppress the oxygenase side reaction ribulose-bisphosphate carboxylase by high CO_2 concentrations. Rapid and strong changes in metabolite pool sizes were induced. The excess of assimilated carbohydrate appeared to be

rapidly sequestered into the increasing sucrose pool which was remobilized during chase. Besides high modulations of many metabolite pool sizes, the fractional enrichment of some metabolites indicated utilization and turn-over of pre-assimilated internal carbon stores, possibly derived from glycogen, lipid or protein break down. Following fast initial kinetics, ~10 min after the pulse, indications of ^{12}C-utilization became obvious. For example citrate and aspartate exhibited negative ^{13}C-enrichment during pulse and fumarate showed a "quasi steady state" at intermediate fractional enrichment similar to our observations using higher plants (Fig. 13). Clearly, a balance between mobilization of internal carbon and *de novo* CO_2 assimilation must be considered for the future modeling of *Synechocystis* primary metabolism under excess CO_2. Furthermore the pulse and chase kinetics of our pilot study were different (Fig. 13) reflecting the different speed of carbon exchange using fast substitution by carbonate addition compared to centrifugation and slow CO_2 exchange by aeration. Aeration exhibited a lag phase indicated by the delayed chase kinetics of the most rapidly responding 3-phosphoglycerate and phosphoenolpyruvate pools. Future experimental optimization of dynamic flux studies will focus on approximating steady state conditions during the pulse and on avoiding large changes of CO_2 availability. Setting pulse and chase by centrifugation and rapid resuspension into differentially labeled but equally concentrated carbonate media appears to be one promising approach to obtain high fractional enrichments. Alternatively CO_2 isotopomer dilution by a factor of 2-3 and monitoring initial kinetic only may also be pursued.

In conclusion, the basic technology, namely, the GC-TOF-MS based dynamic flux analysis of soluble metabolite pools, has been established in pilot studies using $^{13}CO_2$ labeling of plants, *Arabidopsis thaliana* and *Oryza sativa* (**Huege et al. 2007 [7]**), and the *Synechocystis* sp. PCC 6803 reference strain (Huege J, Hagemann M, Kopka J, unpublished). Thus, the ground braking experimental setup of Calvin and co-workers can now be revisited with modern tools. This option may now enhance our understanding of the dynamics of carbon partitioning in photoautotrophic organisms.

The current state of flux technology development in my laboratory is, however, still incomplete and the cyanobacterial flux experiments clearly need optimization towards improved conditions approximating steady states. Relative changes of metabolite pool size and fractional enrichment can be monitored in parallel using the sum and intensity distribution of representative mass isotopomers in kinetic pulse and chase experiments (Fig. 13). Even though combined analysis is now enabled, it is still unclear if apparent relative changes of pool sizes can be used to correct the fractional isotope enrichment for fluctuations of

metabolite concentrations. The requirement to quantify metabolite concentrations is evident. Only exact knowledge of the pool sizes will allow assessment and comparison of the molar carbon partitioning over time and between different metabolite pools. For this reason the initial experiments establishing linear ranges of quantification (**Fiehn et al. 2000a [1]**, **Roessner et al. 2000 [2])** need to be revisited and quantitative calibration curves to be established.

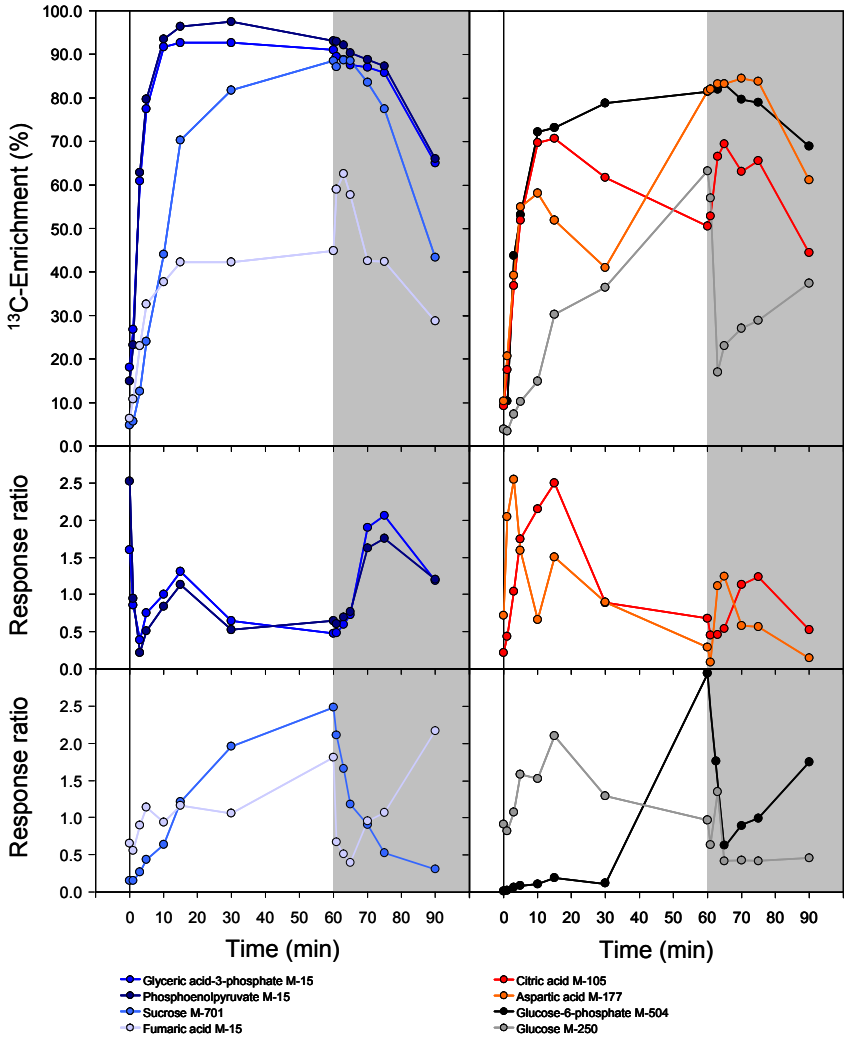


Figure 13 *Pulse and chase kinetics of ^{13}C-isotope dilution into exemplary metabolite pools of a photoautotropic culture of Synechocystis sp. PCC 6803.* A ^{13}C-carbonate pulse (t = 0 min) using growth medium supplemented with excess $NaH^{13}CO_3$ and a chase (t = 60 min) by centrifugation, resuspension in carbonate free medium and ambient aeration was performed. Response ratios representing the relative changes of metabolite pool sizes and ^{13}C-fractional enrichment were calculated. Fluctuations of pool sizes and utilization of non-labeled internal carbon stores during pulse and chase are demonstrated. The influence of internal carbon stores is indicated by intermittent reduction of fractional enrichment during the pulse period, cf. citrate and aspartate. The mass fragments used for this study are identified in the legend by loss of mass units from the molecular ion (M). Note that the response ratios while proportional to the changes of concentration still require exact quantitative calibration. (Huege J, Hagemann, M and Kopka J, unpublished).

The Metabolic Component of Plant Environmental Stress Acclimation

Most, if not all, plant responses to environmental stresses have a metabolic component. Basic stress tolerance as well as acclimation processes and acquired tolerance integrate transcriptional and metabolic components of the plant system. Notably, many of the described abiotic stress effects relate to elements of central metabolism. In this respect, the metabolic responses to abiotic environmental stress conditions differ fundamentally from biotic stress cues which are part of plant-plant, plant-microbe or plant-animal interactions. To manage these interactions plants have evolved an intricate and highly specific exchange of small metabolic signals. These signals enable early detection of pathogen and predator attack or recognition and establishment of beneficial symbioses. In symbiotic and successful parasitic interactions primary metabolites may function as nutrients. However, the signals exchanged in these interactions are often derived from highly specialized, species- or even ecotype-specific secondary metabolic pathways and are typically perceived by specific receptor proteins and signaling cascades. Such secondary metabolites, specifically volatiles and chromophores, are also used for long distance communication. Systemic and local plant metabolic defense responses are typically based on fast evolving deterrent, anti-nutritive and toxic secondary products. Coincident changes of primary metabolism may be interpreted as the obligatory recruitment of resources and energy for facultative secondary metabolite biosynthesis.

In contrast, metabolites involved in abiotic environmental stress responses may have less specific and less obvious functions. Two general rationales of environmentally induced metabolic remodeling are conceivable. On the one hand, owing to the direct effects of physicochemical parameters on catalytic properties of enzymes and membrane transport, environmental cues are discussed to require regulatory adjustments of metabolite steady-state levels and flux in order to re-establish homeostatic conditions (Fernie et al. 2005, Schwender et al. 2004). Metabolic homeostasis may become distorted for example by general thermodynamic effects, such as the temperature dependency of chemical reaction rates, or by an environmental influence on single reactions which can be caused by stress sensitive enzymes. Such mechanisms would both involve patterns of changes throughout the metabolome rather than affect single metabolite pools or reaction rates. On the other hand tolerance mechanisms are thought to utilize important molecular properties of specific metabolites or metabolite classes. In the past particular interest was focused on metabolites that can function (i) as osmolytes to adjust cellular water potential to progressive dehydration,

(ii) as compatible solutes that function to stabilize and chaperone enzymes, membranes and other cellular components, (iii) as chelating agents or scavengers to neutralize or sequester potentially toxic levels of inorganic ions, metals and reactive radicals, such as oxygen species, (iv) as buffering agents to maintain ionic balance and pH homeostasis, (v) as energy sources to fuel cellular damage repair or membrane transport processes involved ion exclusion and cellular or systems wide ion partitioning, and (vi) as reagents which function in re-tailoring the liquid/crystalline physical raft structure of biological membranes. Specific focus has been given to soluble sugars, polyols, polyamines, organic and amino acids as well as lipids (e.g. Guy 1990, Levitt 1972, Thomashow 1999).

In conclusion, the GC-MS based metabolite profiling technology ideally covers - with the exception of complex lipids - the relevant metabolite species of environmental stress responses. Therefore, a non-targeted, within technological limits comprehensive, assessment of metabolic remodeling and underlying patterns in higher plants was initiated. These studies focused on the attempt to correlate changes of metabolite pool sizes to temperature and salt stress cues. In the following the central results and discoveries of these functional metabolic studies will be exemplified. Finally, the studies on the temperature and salt acclimation processes will be used to discuss potential general properties of metabolic stress acclimation.

The Temperature Stress Response of *Arabidopsis thaliana*: A Time Course Study

Time course experiments are perhaps the most informative experimental designs in plant molecular physiology. The sequence of systems responses to experimental interventions can be the basis for hypotheses which try to explain the causal sequence of observed events. In this aspect time course experiments, even though experimentally highly demanding, reveal aspects that can not be observed in purely correlative experimental designs (e.g. Steuer et al. 2003, Urbanczyk-Wochniak et al. 2003, Weckwerth et al. 2004b). Coincident correlative observations define an interaction between systems elements such as metabolite pools and mRNA levels, but usually do not allow deduction of cause and effect relationships. The sequence of events, however, introduces an essential restriction on causal interpretation, because obviously only the preceding events can trigger subsequent effects. To explore the potential of such approaches my laboratory performed - in co-operation with an expert laboratory on plant temperature stress (Guy CL, University of Florida, Gainsville, USA) - the perhaps first comparative time course study of the metabolic responses to long term cold acclimation versus short term heat shock in *Arabidopsis thaliana* Col-0 leaf. Paired

metabolome and transcriptome data were recorded using - at that time feasible - low resolution time courses, namely, 0, 1, 4, 12, 24, 48 and 96h cold acclimation to 4°C and 0, 5, 15, 30, 60, 120 and 240 min heat shock at 40°C (**Kaplan et al. 2004 [17]**, **Kaplan et al. 2007 [18]**). The central result of this study was surprising. Cold acclimation influenced metabolism far more profoundly than heat shock (cf. Fig. 9). As these studies have recently been reviewed in detail (Guy et al. 2008b) four selected aspects will be discussed in the following.

Enzyme encoding transcript abundance compared to metabolite levels. Three types of response relationships between these two systems levels were revealed: (i) Increases in transcript levels for enzymes of defined pathways such as the raffinose biosynthesis and the γ-aminobutyric acid (GABA) shunt preceded increases in metabolite levels consistent with potential regulation by transcriptional activation. (ii) Within some biosynthetic pathways, including glycine, alanine, threonine, leucine and sucrose synthesis or the citric acid cycle the increase of metabolite levels preceded increases in transcript abundance indicative of a possible feedback regulation acting at transcript level. (iii) Other pathways, such as those for lysine, methionine, tryphtophan, tyrosine, arginine, cysteine, as well as polyamine and phenylpropanoid biosynthesis, exhibited decreasing transcript levels for many genes, while the corresponding metabolite pool sizes increased suggesting inverse transcriptional regulation by inactivation of degrading or consummation processes. In conclusion, comparative transcript and metabolome studies can indicate different modes of transcriptional regulation, but may also point towards modes of regulation at other systems levels, namely, in cases of altered metabolite levels which are not accompanied by coincident observations at transcriptome level. In all cases profiling studies lead to hypotheses which require detailed follow-up investigations. These studies are perhaps best performed by reverse genetic approaches. Specifically the dissection of the causal role of signaling components in acclimation events will become highly revealing. For example, the constitutive over-expression of CBF3, the central transcriptional regulator of cold acclimation, demonstrated that ~90%, but not all metabolite pools showing cold acclimation responses, were changed (Cook et al. 2004). Cook and co-authors concluded that the CBF3 cold response pathway plays a prominent but non-exclusive role in metabolic cold acclimation. Especially, the accumulation of fructose, glucose, galactinol, raffinose, proline and 9 yet non-identified MSTs appeared to be downstream of CBF3 signaling (Cook et al. 2004). While a first step towards understanding metabolic regulation under cold stress has been taken, a detailed mechanistic analysis of all signal transduction events between the CBF3 transcription factor

and the change of a defined metabolite pool size is still missing. Indeed CBF3 will represent only one of the multiple signals that may be utilized for the integration of environmental cues at metabolic level.

The raffinose pathway. Many of the temperature stress responses were shared between cold acclimation and heat shock, such as the changes of citric acid cycle intermediates, aromatic amino acids and amino acids levels derived from pyruvate or oxaloacetate. The perhaps most striking similarity between both temperature cues was the reconfiguration of carbohydrate metabolism towards galactinol and raffinose biosynthesis (Guy et al. 2008b). The observed transcriptional activation appeared to involve temperature dependent differential expression of some of the four raffinose synthase and three galactinol synthase isozymes (Guy et al. 2008b). This differential transcript response indicated an intricate mode of regulation for the raffinose pathway in addition to the CBF3 mediated signal input.

The induction of the raffinose pathway was a previously known response to cold acclimation. Chemical properties of galactinol and raffinose were hypothesized to represent a causal mechanism of cold tolerance. Surprisingly, the same pathway also responded to heat shock (**Kaplan et al. 2004 [17]**) and was, therefore, also implicated in the acquisition of heat tolerance (Panikulangara et al. 2004). However, the targeted genetic modification of galactinol and raffinose levels by knock-out and over-expression of galactinol synthase or raffinose synthase genes caused substantially changed metabolite pools but failed to demonstrate significantly altered plant performance under both, heat (Panikulangara et al. 2004) and cold stress (Zuther et al. 2004, Hincha et al. 2006). Thus, the role of raffinose pathway induction under temperature stress still remains elusive. Additional observations, namely, the dose dependent accumulation of galactinol and raffinose in the course of *Arabidopsis thaliana* Col-0 acclimation to NaCl stress (cf. below, **Sanchez et al. 2008b [21]**) or to high light (Kopka J, unpublished) further increased the known instances of raffinose pathway induction in *Arabidopsis thaliana*. Considering all these observations, the difficulty of dissecting the role of single metabolites or enzymes for specific environmental stresses becomes apparent.

Starch mobilization. The process of starch mobilization under cold stress demonstrated the importance of kinetic aspects for the understanding of acclimation responses to environmental stress. Specifically the speed and timing of such responses should be highly informative. During cold acclimation transient increases of maltose and maltotriose pools were discovered which went unnoticed in studies comparing steady states after long-term acclimation. The early maltose transient was followed by a sequence of subsequent increases

in many central carbohydrate pools, e.g. glucose, fructose, glucose-6-phosphate, fructose-6-phosphate, as well as sucrose. Finally, the carbohydrate reallocation extended to late galactinol, raffinose and melibiose accumulations. A non-linear principal component analysis (PCA) corroborated the importance of early maltose accumulation among the observed overall non-linear metabolite patterns (**Scholz et al. 2005 [12]**). This carbohydrate redistribution process appeared to be pushed by rapid starch degradation and was accompanied by transcriptional activation of the disproportionating enzyme and the α-glucan-phosphorylase (Guy et al. 2008b). Furthermore, the plastid targeted β-amylases, BMY7 and BMY8, were induced and maltose was found to have properties of a compatible solute (Kaplan and Guy 2004). Thus, maltose was suggested to play a protective role in acute temperature stress responses (Kaplan and Guy 2004). Targeted genetic modification of β-amylase, BMY8, confirmed a specific protective role of fast maltose or maltose dependent carbohydrate accumulation under cold shock, possibly acting through influencing the photochemical efficiency of photosystem II (Kaplan and Guy 2005). The role of maltose may currently be understood as the fastest available resource for a compatible solute. This function may later be taken by other metabolites with similar solute properties. In other words *Arabidopsis thaliana* may use different but equally effective compatible solutes according to the kinetic and metabolic constraints which are concomitant to the respective stress cues at hand.

Salicylic acid signaling. The importance of metabolite signals for environmental stress responses may best be exemplified by following the kinetics of a typical phytohormone, salicylic acid (SA). The SA levels exhibited - within the first 5 min - a rapid transient increase in response to heat shock and also became elevated during prolonged exposure to low temperature, however, on a much slower time scale (**Kaplan et al. 2004 [17]**, Guy et al. 2008b). The responses of SA levels to both temperature stresses suggest that at least in one aspect, phytohormone signaling may converge during heat and cold stress (Fujita et al. 2006). Several further studies investigated SA signaling and temperature stress. At low temperatures SA levels were found to modulate plant growth, acting under conditions in which the growth rate inversely correlated with freezing tolerance (Scott et al. 2004). Thus, SA signaling has become implied in both, basic (Clarke et al. 2004) and acquired (Larkindale et al. 2005) thermo-tolerance.

In conclusion, the time course study revealed important general aspects of the metabolic complement of environmental stress acclimation: (i) Identical metabolites can be recruited by *Arabidopsis thaliana* under diverse stress conditions. This observation indicates

that it will be hard to assign clear cut functions to a single metabolite. Instead, metabolite patterns may become important diagnostic markers. (ii) Both, the short term and long term kinetics of metabolic stress acclimation, are important. It is noteworthy that cold acclimation does not reach a steady state even after 4 days at 4°C (**Kaplan et al. 2004 [17]**) or longer exposures (Klotke et al. 2004). (iii) The function of primary metabolites may reside in both, the specific chemical properties, e.g. as a compatible solute, and in the provision of sufficiently high metabolite concentrations for biosynthetic or energy conversions. Considering these general functions it is easily conceivable that a plant may choose between alternative metabolites serving the same purpose depending on the coincident environmental, nutritional and metabolic conditions. Two speculative examples, which have not yet been investigated in detail, may serve to illustrate this aspect. It is reasonable to assume that the fast maltose transient under cold stress will not be operative under carbon starvation conditions and needs to be substituted, for example after an extended night or at first light when the levels of transitory starch are low. Also nitrogen limiting conditions may require metabolic plasticity in response to stress, because nitrogen containing solutes, such as proline, may not be available for accumulation to sufficiently high concentrations. In both cases a successful metabolic response should have the option of alternative metabolite usage. (iv) The dynamic change of metabolite patterns will reflect the reallocation of metabolic building blocks and resources and may also comprise signals. (v) Observations of relevant metabolite patterns will not necessarily follow linear interactions or correlations. Instead, non-linear patterns may represent predictive indicators of the plant capacity to tolerate environmental stress. It may, therefore, not come as a surprise that the magnitude of the global metabolome changes which were observed in 9 differentially adapted *Arabidopsis thaliana* ecotypes did not correlate with the acquisition of freezing tolerance during cold acclimation (Hannah et al. 2006). Even though previous studies indicated a differential metabolic response of single ecotypes (e.g. Cook et al. 2004, Klotke et al. 2004) the above studies neglected the kinetic aspect of acclimation and may have overlooked important non-linear and intermediate transient responses.

The essential lesson to be learned from this study relates to the use of linear correlative approaches to data integration. These techniques have been successful in my own and other laboratories for the mining of high-throughput transcription profiling compendia (e.g. **Steinhauser et al. 2004a [15]**, **Steinhauser et al. 2004b [16]**, Lisso et al. 2005, Rautengarten et al. 2005) and can also be used for the analysis and integration of metabolome and transcriptome data (Urbanczyk-Wochniak et al. 2003, Weckwerth et al. 2004b). However,

given the restrictions of traditional experimentation in plant physiology, which mostly relies on pair-wise or multiple comparisons and low resolution time courses, these techniques will probably not yield direct causal hypotheses on regulatory interactions between metabolic and transcriptional systems levels. Novel experimental designs, for example time shifted correlation analyses of high resolution time course experiments, are clearly required for the enhanced and deeper understanding of plant stress physiology.

The Salt Stress Response of *Lotus japonicus*: A Study on Stress Dosage

Parallel to the investigations of temperature stress the analysis of the metabolic complement of salinity stress was initiated in my laboratory as one of the crucial factors affecting global crop productivity. Water limitation is probably the most important environmental constraint. Given (i) the ever-increasing demand for food and animal feed as the world population rises and (ii) the limitation to expand rain-fed crop production, intensive irrigation will be a key strategy in developed and developing countries to meet the global agricultural need. Hence, secondary soil salinization will increasingly affect world agriculture, especially in the expanding arid and semi-arid regions. Unfortunately, we are ill prepared to meet the challenge of soil salinization because most traditional crops are salt sensitive. My laboratory started to explore the metabolic basis of salt acclimation using three model species, the standard plant molecular physiology model, *Arabidopsis thaliana*, the feed plant, *Lotus japonicus* (LOTASSA project, EU INCO-CT-2005-517617, in co-operation with the expert laboratory of Dr. MK Udvardi, MPIMP), and the monocot rice crop, *Oryza sativa* (cf. Zuther et al. 2007).

Comparative analysis of the metabolic salt stress response. As a systematic investigation of model plants and crop species is clearly required, the initial work was focused on a comparative analysis of metabolic acclimation processes of the different species, so as to estimate the potential to transfer knowledge from models to crops and, thus, the possibility to generalize biotechnological approaches towards modified salinity tolerance. Non-lethal dosage dependency analyses were performed using increasing saline conditions and long term acclimation (> 14 days). Experiments were performed with 4-6 replicate plants of each condition and in three independent repeats to eliminate spurious responses and biases that may occur in a single experiment (cf. Fig. 1A of **Sanchez et al. 2008b [21]**). The early reports on the response of higher plants to salt stress, but also the new profiling results (e.g. Avelange-Macherel et al. 2006, Cramer et al. 2007, Kim et al. 2007, Pinheiro et al. 2004,

Rizhsky et al. 2004), appear at first sight to yield variable results. This may be explained in part by inherent differences between plant species, but may also be related to the use of, different organs, time scales of stress exposure, modes of cultivation, e.g. soil and hydroponic growth, and difficult to control combinations of stress factors, such as osmotic, drought, salt toxicity, oxidative stress and nutritional conditions.

Only few metabolites exhibited conserved responses among the three investigated species. Essentially all of the conserved responses were found in central metabolism. The central carbohydrates glucose, fructose, glucose-6-phosphate and fructose-6-phosphate were depleted in leaf tissue. The balance between organic acid and amino acid levels was also changed. In general organic acids decreased whereas amino acid levels increased. Specifically citrate, 2-oxoglutarate, succinate and malate were strongly reduced in all three species and most other acids were reduced at least in two of three species. In contrast proline, glycine and serine appeared to be consistently increased, while leucine, isoleucine and valine accumulated in at least two species.

In contrast to the few conserved responses a rich diversity of divergent metabolic acclimation reactions was observed. For example, raffinose and galactinol pools accumulated in *Arabidopsis thaliana* and *Oryza sativa* leaf tissue but were depleted in *Oryza sativa* roots. In contrast, *Lotus japonicus* appeared to utilize *myo*-inositol an intermediate of raffinose biosynthesis for the synthesis of methyl-inositols, such as ononitol and pinitol. These compounds are frequently found in legumes and discussed to represent compatible solutes in some halophytes (Nelson et al. 1998). A second example of divergent metabolic responses to salt stress was found among tryptophan related metabolites. Tryptophan, 5-hydroxy-tryptamine and to a minor degree the intermediate tryptamine pool accumulated in rice leaf tissue whereas rice roots exhibited an inverse behavior. This response was not observable in *Lotus japonicus* or *Arabidopsis thaliana*. (cf. Fig. 6 of **Sanchez et al. 2008b [21]**).

We concluded that the detailed analysis of the effects of salt acclimation and related stress factors, such as osmotic stress and drought, in time and dose may help to direct breeding programs toward increased salt tolerance, but should best be performed with focus on a specific crop species. In the following the results on *Lotus japonicus* are shortly discussed as these investigations are currently the most detailed contributions of my laboratory to the salt stress physiology field.

The salt stress response of Lotus japonicus. In co-operation with the expert laboratory of Dr. MK Udvardi *Lotus japonicus* was initially established as a model system for legume metabolite profiling (**Desbrosses et al. 2005a [20]**, Desbrosses et al. 2005b). The combined

metabolic and transcriptomic analysis of physiological phenomena proved feasible (**Colebatch et al. 2004 [19]**) and the metabolite profiling application was readily transferred to different legume species, e.g. *Medicago truncatula* (Lohse et al. 2005, Schaarschmidt et al. 2007) and *Phaseolus vulgaris* (Hernandez et al. 2007, Hannah et al. 2007). These initial studies demonstrated the high metabolic diversity even between different organs of the same plant and between different parts of the same organ, such as primary and secondary roots or nodules (**Colebatch et al. 2004 [19]**), or developmental stages, e.g. developing and mature leaves (**Desbrosses et al. 2005a [20]**).

As the number of factors which may influence the systems response to salt stress dosage was high and hard to control we decided to broadly monitor the combinatorial interactions of systems responses at the levels of plant growth, ion balance and mineral nutrition as well as metabolism and transcription again using three replicate experiments of long-term (28 days) salt acclimated plants (**Sanchez et al. 2008a [22]**). Coordinated salt stress-induced reduction of growth as well as decreasing leaf potassium, phosphorus, sulphur, zinc and molybdenum levels were described. This systems wide investigation on 912 salt responsive transcripts and 147 differential metabolic components appeared to be in agreement with the established biphasic model of salt response (Munns 2002, Munns 2005). Results were largely independent of the mode of salt application, namely, either step-wise increase of salt dosage or exposure to an initial and constant salt concentration. Current data indicate similarity of early changes with osmotic or drought responses while long-term salt acclimation may reflect the attempt to cope with increasing ion toxicity finally reaching a critical state dependent on dose and time of salt exposure. Similar to the time course responses in temperature acclimation, we found evidence of non-linear stress responses and thus the necessity to modify or fine-tune the initially proposed strictly linear qualitative dose-dependency model. We demonstrated threshold behavior at both, low and high salt concentrations, but found no apparent transient dosage dependent maxima or minima of metabolite pools using 5-6 different treatments from 0 to 150 mM NaCl.

In agreement with our observations from temperature induced time course experiments we found a successive and increasingly global requirement for the reprogramming of gene expression and metabolic pathways rather than rapid transitions between clearly defined states. Moreover the individual variation especially of the metabolic phenotype increased comparing highly repeatable control individuals to the increasingly variable salt stressed individuals (cf. Fig. 6A of **Sanchez et al. 2008a [22]**). Finally, following the findings of Meyer and co-authors (2007) we investigated the predictive power of metabolome based

models for the biomass and ionome parameters after salt acclimation (**Sanchez et al. 2008c [23]**). Using an alternative statistical multivariate regression technique called orthogonal projections to latent structures (OPLS; Trygg et al. 2002, Bylesjö et al. 2007) we demonstrated excellent correlation of the metabolome and the ionome and biomass which may allow the estimation of the degree of salt stress experienced by a plant based on metabolite profiles taken under these conditions. Despite the apparently high predictive power of the OPLS models using simultaneously recorded data of metabolite changes, ion levels and biomass, it remains to be investigated whether metabolite profiles taken from non- or moderately-stressed plants can be used to predict the later outcome and success of acclimation and salt stress survival. Only truly predictive systems features may have the potential to serve breeding programs as ideal ideotypes for the selection of enhanced salt-tolerant genotypes.

In conclusion, it may be possible in the future to use metabolic fingerprinting as a breeding tool to select individual plants that best cope with salt stress. But further experimentation focusing on truly predictive systems features representing pre-formed genetically encoded adaptations is required. Also accessing the natural variation of salt tolerance within the *Lotus japonicus* species, for example by the QTL mapping approach which proved to be highly successful in analyses of tomato fruit traits (e.g. Schauer et al. 2006), will lead towards discovery of relevant functional genes and metabolites.

On the other hand, given the interdependent nature of plant responses to diverse environmental and nutritional stresses, metabolite-based models may not reveal unique predictors of modified growth under a single stress factor. A deeper comparison of the salt stress response to related stress factors, such as osmotic and drought stresses, is indicated. Due to the high diversification of biosynthetic capabilities, the transfer of knowledge and principles between species even belonging to closely related plant clades may be restricted and further investigation for example of closely related legume species in comparison to ecotypes of a single species may provide more detailed insights for a rational design of novel reverse transgenic approaches or smart selection for enhanced plant breeding.

SUMMARY AND PERSPECTIVES

The central objective of this thesis is the establishment, application and enhancement of gas chromatography-mass spectrometry (GC-MS) based metabolite profiling as a routine phenotyping tool box in plant biosciences by interfacing elements of analytical chemistry, bioinformatics and molecular physiology. In response to the demand for standardized procedures in the emergent metabolomics field, I aimed my efforts at enabling the transfer of techniques and results between laboratories and at the development of software-supported workflows for improved and reproducible chemical analysis of metabolite profiling experiments, including enhanced chromatography data processing, compound identification and physiological interpretation. All projects towards the improvement of technological capabilities were fine-tuned to the needs of plant molecular physiology using key cooperation partners. Thus, I have established an applied focus on metabolic aspects of plant environmental stress acclimation for my work which motivates my technological aims.

After my initial contributions to the establishment of GC-MS based metabolite profiling in both, academic and industrial environments, starting in 1997 and 1998, respectively, the major technological achievements were (i) an enhanced laboratory automation by miniaturized in-line chemical derivatization and automated, high-throughput GC-time of flight (TOF)-MS, (ii) the development of a software-supported, standardized procedure implemented in the TagFinder package for GC-TOF-MS chromatography data processing of and metabolite recognition within large scale experiments, (iii) the initiation of the Golm Metabolome Database (GMD), (iv) the development of enhanced precision metabolite profiling using mass isotopomer ratios (ITR) of *in vivo*-labeled whole organisms and (v) the extension of GC-MS profiling technology towards combined quantification of pool sizes and flux.

The most fundamental discovery of this work was the high number of yet non-identified metabolites within GC-TOF-MS profiling experiments which may amount to ~1000 observable metabolic components. This basic finding of unexpectedly high metabolic complexity was later found to apply to essentially all analytical technology platforms of modern metabolomics. Thus, the identification of the complete metabolic complement of biological systems may represent the grand challenge of the metabolomics field tackled in this thesis starting with the establishment of an international mass spectral and retention index library. These data were the first to be made publicly available by the GMD which since has been continuously extended by reference data of authenticated pure metabolite preparations

and by novel, yet non-identified, mass spectral tags (MSTs). The rich resource of reference data has been utilized for TagFinder-supported compound identification and structural interpretation of novel MSTs. The development of metabolite profiling using metabolite mass isotopomer ratios (ITR) of *in vivo*-labeled whole organisms, such as the yeast *Saccharomyces cerevisiae*, as multiplexed quantitative internal standards, has not only demonstrated the path towards high precision metabolite profiling but also provided stable isotope labeled mass spectra to support the MST identification process. Moreover, the grounds were prepared for the development towards the combination of quantitative pool size and flux analysis. The combination of pool size and flux analysis is still in the early stages of technology development, but already the discovery has been made that higher plants, in contrast to cyanobacteria, do not rapidly approach full labeling of primary metabolite pools. This finding will contribute to our understanding of how higher plants may organize the use of *de novo* assimilated carbon resources compared to the previously assimilated internal carbon stores.

The continuous application of the GC-TOF-MS tool box to plant molecular physiology has lead to novel insights into the metabolic aspects of plant stress acclimation. The global knowledge gained from studies on environmental stress acclimation to temperature and salt cues may best be characterized by the finding that higher plants do not appear to reach distinct steady states of metabolite pools. In contrast metabolite pools seem to continuously adjust to the immediate progressive needs. The response modes, both, to the time of exposure to temperature stress and to the dosage of salt stress, were non-linear. Moreover, metabolite profiles have been shown to have a high predictive power for both, the changed ion levels and the biomass production in response to salt stress.

The long-term aim of this thesis has been the development of a data resource within the framework of GMD which combines both, the qualitative aspects of metabolite identification and large, well described compendia of GC-MS based metabolite profiles for functional analysis. GMD is envisaged to become a unique resource for metabolic reference data. The enhancement of non-targeted GC-MS profiling technology has been tuned towards reproducible analytical performance and standardized data processing which were prerequisite for the creation of a database. Currently, GMD harbors cheminformatic information on metabolite identity and structures which are relevant for GC-MS based profiling. The uploading of metabolite profiles which describe the changed metabolite pools in relation to biological and chemical references has begun. Information on pre-exisiting and new experiments on genetic and environmental factors which influence the metabolome will be accumulated and made available for comparative physiological analysis. As the metabolome

appears to adopt variable states influenced by multiple factors, the future data mining will require bioinformatic machine learning support. Ultimately, this work intends to enhance our understanding of the metabolic systems complement and to enable the discovery and elucidation of new, potentially predictive, metabolic markers and their role in environmental stress acclimation.

ACKNOWLEDGEMENTS

The work presented in this thesis would not have been possible without the support of many individuals, to whom I am deeply grateful and without the funding by several organizations which enabled my research approach at the interface of analytical chemistry, bioinformatics and molecular plant physiology.

I dedicate this thesis to my family. My parents, Albertine Kopka, né André, and Manfred Kopka, recognized my interest in natural sciences, encouraged me to complete my school and university education and gave substantial emotional, intellectual and financial support. My own family has stayed at my side and is the safe haven for my life. I deeply acknowledge that my wife, Andrea Kopka, né Scholthaus, dedicated part of her life to me. My children, Johanna and Jonas, grew up with a father engaged in academic science and have formed an understanding of the inevitable time demands and necessities implied. I hope that my family may find similar support from me.

I thank most of all Prof. Dr. Lothar Willmitzer for his mentoring over the last decade, his unyielding support for this work and the role he has played facilitating my scientific achievements at the Max-Planck-Institute of Molecular Plant Physiology.
I thank Prof. Dr. Tiedemann, Prof. Dr. Bernd Waltz, Prof. Dr. M. Steup and the "Institutsrat" of the Institute of Biochemistry and Biology at the University of Potsdam for the possibility of experiencing and performing university teaching, for their advice and for their decision to support my "Habilitationsanliegen".
My gratitude extends to the directors of the Max-Planck-Institute of Molecular Plant Physiology, Prof. Dr. Lothar Willmitzer, Prof. Dr. Mark Stitt and Prof. Dr. Ralph Bock, for providing an excellent and interactive environment for plant scientists. Moreover, the help of the administration, the workshop and greenhouse teams at the Max-Planck-Institute of Molecular Plant Physiology is unsurpassed. Most of all, I value the stimulating discussions and fruitful co-operations with many members of the Max-Planck-Institute of Molecular Plant Physiology. These interactions were essential to my work. I would like to specifically acknowledge Dr. Alisdair R. Fernie, Dr. Oliver Fiehn, Dr. Ute Roessner, Prof. Dr. Wolfram Weckwerth, Prof. Dr. Peter Dörmann, Prof. Dr. Thomas Altmann, Dr. Joost T. van Dongen, Dr. Peter Geigenberger, Dr. Arnd Heyer, Dr. Dirk Hincha, Dr. Rainer Höfgen, Dr. Joachim Fishan, Dr. Vica Nikiforova, Dr. Ute Krämer, Dr. Yves Gibon, Prof. Dr. Joachim Selbig, Dr. Dirk Walther, Dr. Jan Hummel, Dr. Matthias Scholz, Dr. Matthias Steinfath, Dr. Karin Köhl, Dr. Ellen Zuther, Dr. Christian Bölling, Dr. Stefan Kempa, Dr. Matthew A. Hannah and Dr. Nicholas Schauer.
Outside the Max-Planck-Institute of Molecular Plant Physiology there are many scientists from whom I have learned much. Prof. Dr. Werner Bottke introduced me to experimental biological science. Prof. Dr. Friedrich Spener gave me the opportunity to continue my studies in biochemistry and supervised my PhD thesis. Prof. Jan Jarworsky, Prof. Bernd Müller-Röber and Dr. Richard Trethwey formed my post doctoral career and fuelled my interest in plant physiology. I am deeply grateful to Dr. Richard Trethwey, Dr. Arno Krotzky and Prof. Dr. Lothar Willmitzer for offering me the opportunity to participate in founding the Metanomics GmbH & Co. KGaA and to return to academic science afterwards.
I feel in great debt to all the guest scientists and cooperation partners who stayed in my group. There have been too many to acknowledge them all. The list of co-authored publications, however, vividly reflects these fruitful activities of my laboratory. For the closest scientific interactions I would like to extend my special thanks to Dr. Alisdair R. Fernie, and Dr. Nicholas Schauer, to Dr. Michael K. Udvardi and Dr. Guilhem G. Desbrosses, to Prof. Dr. Charles L. Guy and Dr. Fatma Kaplan and to Prof. Dr. Martin Hageman. These scientists enhanced my deep interest in plant physiology and enthusiastically introduced me to the highly diverse aspects of this field.
Important financial support, which has made this work possible, has been granted by the Max Planck Society, the Metanomics GmbH & Co. KGaA, the BMBF, the DFG, the DAAD and the European Union.
Finally, I would like to thank all the members of my laboratory for their enthusiastic work and excellent performance both, in the laboratory and at the computer: Cornelia Wagner, Alexander Lüdemann, Alexander Erban, Katrin Bieberich, Nicole Gatzke, Ines Fehrle, Ania Kolasa, Dr. Stephan Krüger, Jedrzey Szymanski, Franziska Schwabe, Stefanie Schmidt, Sepideh Bijanzadeh, Mohammad Reza Siahpoosh, Katrin Strassburg, Jan Hüge, Luise von Malotky, Nadine Strehmel, Dr. Dirk Steinhauser, Dr. Claudia Birkemeyer, Dr. Takeshi Shoji, and Dr. Diego H. Sanchez. Without their support the work described in this thesis would not have been possible.

Potsdam-Golm,

REFERENCES

America AHP, Cordewener JHG, van Geffen HA, Lommen A, Vissers JPC, Bino RJ, Hall RD (2006) Alignment and statistical difference analysis of complex peptide data sets generated by multidimensional liquid chromatography mass spectrometry. Proteomics 6: 641-653

Ardenkjaer-Larsen JH, Fridlund B, Gram A, Hansson G, Hansson L, Lerche MH, Servin R, Thaning M, Golman K (2003) Increase in signal-to-noise ratio of > 10,000 times in liquid-state NMR. Proc Natl Acad Sci USA 100: 10158-10163

Arita M (2004) Computational resources for metabolomics. Briefings Funct Genomics Proteomics 3: 84-93

Ausloos P, Clifton CL, Lias SG, Mikaya AI, Stein SE, Tchekhovskoi DV, Sparkman OD, Zaikin V, Zhu D (1999) The critical evaluation of a comprehensive mass spectral library. J Am Soc Mass Spectrom 10: 287-299

Avelange-Macherel MH, Ly-Vu B, Delauna Y, Richomme P, Leprince O (2006) NMR metabolite profiling analysis reveals changes in phospholipid metabolism associated with the re-establishment of desiccation tolerance upon osmotic stress in germinated radicles of cucumber. Plant Cell Environ 29: 471–482

Bader GD, Cary MP, Sander C (2006) Pathguide: a pathway resource list. Nucleic Acids Res 34: D504-506

Baran R, Kochi H, Saito N, Suematsu M, Soga T, Nishioka T, Robert M, Tomita M (2006) MathDAMP: a package for differential analysis of metabolite profiles. BMC Bioinformatics 7: 530

Barsch A, Patschkowski T, Niehaus K (2004) Comprehensive metabolite profiling of *Sinorhizobium meliloti* using gas chromatography-mass spectrometry. Funct Integrat Genomics 4: 219-230

Baxter CJ, Redestig H, Schauer N, Repsilber D, Patil KR, Nielsen J, Selbig J, Liu JL, Fernie AR, Sweetlove LJ (2007) The metabolic response of heterotrophic Arabidopsis cells to oxidative stress. Plant Physiol 143: 312-325

Ben Wagner A (2006) SciFinder Scholar 2006: An empirical analysis of research topic query processing. J Chem Information Modeling 46: 767-774

Bino RJ, Hall RD, Fiehn O, Kopka J, Saito K, Draper J, Nikolau BJ, Mendes P, Roessner-Tunali U, Beale MH, Trethewey RN, Lange BM, Wurtele ES, Sumner LW (2004) Potential of metabolomics as a functional genomics tool. Trends Plant Sci 9: 418-425

Bino RJ, de Vos CHR, Lieberman M, Hall RD, Bovy A, Jonker HH, Tikunov Y, Lommen A, Moco S, Levin I (2005) The light-hyperresponsive high pigment-2 mutation of tomato: alterations in the fruit metabolome. New Phytol 166: 427-438

Birkemeyer C, Kolasa A, Kopka J (2003) Comprehensive chemical derivatization for gas chromatography-mass spectrometry-based multi-targeted profiling of the major phytohormones. J Chromatogr A 993: 89-102

Birkemeyer C, Luedemann A, Wagner C, Erban A, Kopka J (2005) Metabolome analysis: the potential of *in vivo*-labeling with stable isotopes for metabolite profiling. Trends Biotechnol 23: 28-33

Blau K, Halket JM (1993) Handbook of derivatives for chromatography, 2nd edn. Wiley Publishers, New York

Böttcher C, Centeno D, Freitag J, Höfgen R, Köhl K, Kopka J, Kroymann J, Matros A, Mock H-P, Neumann S, Pfalz M, von Roepenack-Lahaye E, Schauer N, Trenkamp S, Zubriggen M, Fernie AR (2008) Teaching (and learning from) metabolomics: The 2006 PlantMetaNet ETNA metabolomics research school. Physiol Plant 132: 136-149

Broeckling CD, Huhman DV, Farag MA, Smith JT, May GD, Mendes P, Dixon RA, Sumner LW (2005) Metabolic profiling of *Medicago truncatula* cell cultures reveals the effects of biotic and abiotic elicitors on metabolism. J Exp Bot 56: 323-336

Bunk B, Kucklick M, Jonas R, Muench R, Schobert M, Jahn D, Hiller K (2006) MetaQuant: a tool for the automatic quantification of GC/MS-based metabolome data. Bioinformatics 22: 2962-2965

Bylesjö M, Eriksson D, Kusano M, Moritz T, Trygg J (2007) Data integration in plant biology: the O2PLS method for combined modelling of transcript and metabolite data. Plant J 52: 1181-91

Calvin M (1956) The photosynthetic carbon cycle. J Chem Soc 1895-1915

Calvin M (1962) The path of carbon in photosynthesis. Science 135: 879-889

Calvin M (1964) The path of carbon in photosynthesis. In: Nobel Lectures Chemistry 1942-1962. Elsevier Publishing Company, Amsterdam, pp 618-644

Cary MP, Bader GD, Sander C (2005) Pathway information for systems biology. FEBS Letters 579: 1815-1820

Caspi R, Foerster H, Fulcher CA, Hopkinson R, Ingraham J, Kaipa P, Krummenacker M, Paley S, Pick J, Rhee SY, Tissier C, Zhang PF, Karp PD (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. Nucleic Acids Res 34: D511-516

Castle AL, Fiehn O, Kaddurah-Daouk R, Lindon JC (2006) Metabolomics Standards Workshop and the development of international standards for reporting metabolomics experimental results. Briefings Bioinformatics 7: 159-165

Clarke SM, Mur LA, Wood JE, Scott IM (2004) Salicylic acid dependent signaling promotes basal thermotolerance but is not essential for acquired thermotolerance in *Arabidopsis thaliana*. Plant J 38: 432–447

Colebatch G, Desbrosses GG, Ott T, Krusell L, Montanari O, Kloska S, Kopka J, Udvardi MK (2004) Global changes in transcription orchestrate metabolic differentiation during symbiotic nitrogen fixation in Lotus japonicus. Plant Journal 39: 487-512

Cook D, Fowler S, Fiehn O, Thomashow MF (2004) A prominent role for the CBF cold response pathway in configuring the low-temperature metabolome of *Arabidopsis*. Proc Natl Acad Sci USA 101:15243-15248

Cramer GR, Ergul A, Grimplet J, Tillett RL, Tattersall EAR, Bohlman MC, Vincent D, Sonderegger J, Evans J, Osborne C, Quilici D, Schlauch KA, Schooley DA, Cushman JC (2007) Water and salinity stress in grapevines: early and late changes in transcript and metabolite profiles. Funct Integr Genomics 7: 111–134

Dalluge J, Vreuls RJJ, Beens J, Brinkman UAT (2002a) Optimization and characterization of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometric detection (GCxGC-TOF MS). J Sep Sci 25: 201-214

Dalluge J, van Rijn M, Beens J, Vreuls RJJ, Brinkman UAT (2002b) Comprehensive two-dimensional gas chromatograpy with time-of-flight mass spectrometric detection applied to the determination of pesticides in food extracts. J Chromatogr A 965: 207-217

Desbrosses GG, Kopka J, Udvardi MK (2005a) *Lotus japonicus* metabolic profiling. Development of gas chromatography-mass spectrometry resources for the study of plant-microbe interactions. Plant Physiology 137: 1302-1318

Desbrosses GG, Steinhauser D, Kopka J, Udvardi MK (2005b) Metabolome analysis using GC-MS. In: Marquez AJ (ed) *Lotus japonicus* Handbook, Springer-Verlag, Dordrecht, pp 165-174

De Souza DP, Saunders EC, McConville MJ, Likic VA (2006) Progressive peak clustering in GC-MS metabolomic experiments applied to *Leishmania* parasites. Bioinformatics 22: 1391-1396

De Vos RCH, Moco S, Lommen A, Keurentjes JJB, Bino RJ, Hall RD (2007) Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. Nature Protocols 2: 778-791

Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA 104: 1777-1782

Dueck TA, de Visser R, Poorter H, Persijn S, Gorissen A, de Visser W, Schapendonk A, Verhagen J, Snel J, Harren FJM, Ngai AKY, Verstappen F, Bouwmeester H, Voesenek LACJ, van der Werf A (2007) No evidence for substantial aerobic methane emission by terrestrial plants: a C-13-labelling approach. New Phytologist 175: 29-35

Duran AL, Yang J, Wang LJ, Sumner LW (2003) Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). Bioinformatics 19: 2283-2293

Eisenhut M, Huege J, Schwarz D, Bauwe H, Kopka J, Hagemann M (2008) Metabolome phenotyping of inorganic carbon limitation in cells of the wild type and photorespiratory mutants of the cyanobacterium *Synechocystis sp.* strain PCC 6803. Plant Physiology (First published on October 22, 2008; 10.1104/pp.108.129403)

Engelsberger WR, Erban A, Kopka J, Schulze WX (2006) Metabolic labeling of plant cell cultures with $K^{15}NO_3$ as a tool for quantitative analysis of proteins and metabolites. Plant Methods 2: 14

Erban A, Schauer N, Fernie AR, Kopka J (2007) Non-supervised construction and application of mass spectral and retention time index libraries from time-of-flight GC-MS metabolite profiles. In: Weckwerth W (ed) Metabolomics: methods and protocols. Humana Press, Totowa, pp 19-38

Fernie AR, Trethewey RN, Krotzky AJ, Willmitzer L (2004) Metabolite profiling: from diagnostics to systems biology. Nat Rev Mol Cell Biol 5: 763-769

Fernie AR, Geigenberger P, Stitt M (2005) Flux an important, but neglected, component of functional genomics. Curr Opin Plant Sci 8: 174-182

Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, Willmitzer L (2000a) Metabolite profiling for plant functional genomics. Nat Biotechnol 18: 1142-1143

Fiehn O, Kopka J, Trethewey RN, Willmitzer L (2000b) Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. Anal Chem 72:3573-3580

Fiehn O (2002) Metabolomics - the link between genotypes and phenotypes. Plant Mol Biol 48: 155-171

Fiehn O, Weckwerth W (2003) Deciphering metabolic networks. Europ J Biochem 270: 579-588

Fiehn O, Wohlgemuth G, Scholz M (2005) Automatic annotation of metabolomic mass spectra by integrating experimental metadata. Proc Lect Notes Bioinformatics 3615: 224-239

Fiehn O, Kind T (2007) Metabolite profiling in blood plasma. In: Weckwerth W (ed) Metabolomics: methods and protocols. Humana Press, Totowa, pp 3-18

Fiehn O, Robertson D, Griffin J, van der Werf M, Nikolau B, Morrison N, Sumner LW, Goodacre R, Hardy NW, Taylor C, Fostel J, Kristal B, Kaddurah-Daouk R, Mendes P, van Ommen B, Lindon JC, Sansone SA (2007) The metabolomics standards initiative (MSI). Metabolomics 3: 175-178

Forster J, Famili I, Fu P, Palsson BO, Nielsen J (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res 13: 244-253

Fujita M, Fujita Y, Noutoshi Y, Takahashi F, Narusaka Y, Yamaguchi-Shinozaki K, Shinozaki K (2006) Crosstalk between abiotic and biotic stress responses: a current view from the points of convergence in the stress signaling networks. Curr Opin Plant Biol 9: 436–442

Galperin MY (2005) The molecular database collection: 2005 update. Nucleic Acids Res 33: D5-D24

Goto S, Nishioka T, Kanehisa M (1998) LIGAND: chemical database for enzyme reactions. Bioinformatics 14: 591-599

Gullberg J, Jonsson P, Nordström A, Sjöström M, Moritz T (2004) Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of *Arabidopsis thaliana* samples in metabolomic studies with gas chromatography/mass spectrometry. Anal Biochem 331: 283-295

Guy CL (1990) Cold acclimation and freezing stress tolerance: role of protein metabolism. Annu Rev Plant Physiol Plant Mol Biol 41: 187–223

Guy CL, Kopka J, Moritz T (2008a) Plant metabolomics coming of age. Physiologica Plantarum 132: 113-116

Guy CL, Kaplan F, Kopka J, Selbig J, Hincha D (2008b) Metabolomics of temperature stress. Physiologica Plantarum 132: 220-235

Halket JM, Przyborowska A, Stein SE, Mallard WG, Down S, Chalmers RA (1999) Deconvolution gas chromatography mass spectrometry of urinary organic acids - potential for pattern recognition and automated identification of metabolic disorders. Rapid Commun Mass Spectrom 13: 279-284

Halket JM, Waterman D, Przyborowska AM, Patel RKP, Fraser PD, Bramley PM (2005) Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. J Exp Bot 56: 219-243

Hall R, Beale M, Fiehn, O, Hardy N, Sumner L, Bino R (2002) Plant metabolomics: The missing link in functional genomics strategies. Plant Cell 14: 1437-1440

Hannah MA, Wiese D, Freund S, Fiehn O, Heyer AG, Hincha DK (2006) Natural genetic variation of freezing tolerance in *Arabidopsis thaliana*. Plant Physiol 142: 98-112

Hannah MA, Kraemer KM, Geffroy V, Kopka J, Blair MW, Erban A, Vallejos CE, Heyer AG, Sanders FET, Millner PA, Pilbeam DJ (2007) Hybrid weakness controlled by the dosage-dependent lethal (DL) gene system in common bean (*Phaseolus vulgaris*) is caused by a shoot-derived inhibitory signal leading to salicylic acid-associated root death. New Phytologist 176: 537-549

Harrigan GG, Goodacre R (eds) (2003) Metabolic profiling: Its role in biomarker discovery and gene function analysis. Kluwer Academic Publishers, London, UK

Hegeman AD, Schulte CF, Cui Q, Lewis IA, Huttlin EL, Eghbalnia H, Harms AC, Ulrich EL, Markley JL, Sussman MR (2007) Stable isotope assisted assignment of elemental compositions for metabolomics. Anal Chem 79: 6912-6921

Hernandez G, Ramirez M, Valdes-Lopez O, Tesfaye M, Graham MA, Czechowski T, Schlereth A, Wandrey M, Erban A, Cheung F, Wu HC, Lara M, Town CD, Kopka J, Udvardi MK, Vance CP (2007) Phosphorus stress in common bean: Root transcript and metabolic responses. Plant Physiology 144: 752-767

Herold MM, Dostler M, Looser R, Walk T, Fegert A, Kluttig M, Lehmann B, Heidemann S, Hennig A, Kopka J (2003a) Method for extraction and analysis of contents of organic materials such as plant tissue. PCT Int. Appl. (2003), 38 pp. WO 2003041835 A1 20030522 CAN 138:398388 AN 2003:396761

Herold MM, Dostler M, Looser R, Walk T, Fegert A, Kluttig M, Lehmann B, Heidemann S, Hennig A, Kopka J (2003b) Single-phase mixtures for extraction of components from organic materials such as plant tissue. PCT Int. Appl. (2003), 38 pp. WO 2003041834 A1 20030522 CAN 138:398387 AN 2003:396760

Herold MM, Christiansen N, Kluttig M, Kopka J, Quedenau J (2006) Mass spectrometry analysis method and system. PCT Int. Appl. (2006), 71 pp. WO 2006082042 A2 20060810 CAN 145:198627 AN 2006:796358

Hincha DK, Zuther E, Hundertmark M, Heyer AG (2006) The role of compatible solutes in plant freezing tolerance: a case study on raffinose. In: Chen THH, Uemura M, Fujikawa S (eds) Cold Hardiness in Plants: Molecular Genetics, Cell Biology and Physiology. CABI Publishing, Wallingford, UK, pp 203-218

Huege J, Sulpice R, Gibon Y, Lisec J, Koehl K, Kopka J (2007) GC-EI-TOF-MS analysis of *in vivo*-carbon-partitioning into soluble metabolite pools of higher plants by monitoring isotope dilution after ($^{13}CO_2$)-labelling. Phytochemistry 68: 2258-2272

Hummel J, Selbig J, Walther D, Kopka J (2008) The Golm Metabolome Database: a Database for GC-MS based Metabolite Profiling. In: Nielsen J, Jewett M (eds) Metabolomics A Powerful Tool in Systems Biology, Topics in Current Genetics, Vol. 18. Springer-Verlag, Berlin Heidelberg New York, pp 75-96

Jacobs A, Lunde C, Bacic A, Tester M, Roessner U (2007) The impact of constitutive heterologous expression of a moss Na^+ transporter on the metabolomes of rice and barley. Metabolomics 3: 307-317

Jellum E, Helland P, Eldjarn L, Markwardt U, Marhofer J (1975) Development of a computer assisted search for anomalous compounds (CASAC). J Chromatogr 112: 573-580

Jellum E (1977) Profiling of human body fluids in healthy and diseased states using gas chromatography and mass spectrometry, with special reference to organic acids. J Chromatrogr B 143: 427-462

Jellum E (1979) Application of mass spectrometry and metabolite profiling to the study of human diseases. Philosophical Transactions of the Royal Society of London Series A: Mathematical Physical and Engineering Sciences 293: 13-19

Jenkins H, Hardy N, Beckmann M, Draper J, Smith AR, Taylor J, Fiehn O, Goodacre R, Bino RJ, Hall R, Kopka J, Lane GA, Lange BM, Liu JR, Mendes P, Nikolau BJ, Oliver SG, Paton NW, Rhee S, Roessner-Tunali U, Saito K, Smedsgaard J, Sumner LW, Wang T, Walsh S, Wurtele ES, Kell DB (2004) A proposed framework for the description of plant metabolomics experiments and their results. Nat Biotechnol 22: 1601-1606

Jonsson P, Gullberg J, Nordström A, Kusano M, Kowalczyk M, Sjöström M, Moritz T (2004) A strategy for identifying differences in large series of metabolomic samples analyzed by GC/MS. Anal Chem 76: 1738-1745

Jonsson P, Johansson AI, Gullberg J, Trygg J, A J, Grung B, Marklund S, Sjostrom M, Antti H, Moritz T (2005) High-throughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses. Anal Chem 77: 5635-5642

Jonsson P, Johansson ES, Wuolikainen A, Lindberg J, Schuppe-Koistinen I, Kusano M, Sjöström M, Trygg J, Moritz T, Antti H (2006) Predictive metabolite profiling applying hierarchical multivariate curve resolution to GC-MS data – a potential tool for multi-parametric diagnosis. J Proteome Res 5: 1407-1414

Kanehisa M (1997) A database for post-genome analysis. Trends Genet 13: 375-376

Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28: 27-30

Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. Nucleic Acids Res 30: 42-46

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34: D354-357

Kaplan F, Guy CL (2004) β-Amylase induction and the protective role of maltose during temperature shock. Plant Physiol 135: 1674–1684

Kaplan F, Guy CL (2005) RNA interference of *Arabidopsis* β-amylase-8 prevents maltose accumulation upon cold shock and increases sensitivity of PSII photochemical efficiency to freezing stress. Plant J 44: 730-743

Kaplan F, Kopka J, Haskell DW, Zhao W, Schiller KC, Gatzke N, Sung DY, Guy CL (2004) Exploring the temperature-stress metabolome of Arabidopsis. Plant Physiol 136: 4159-4168

Kaplan F, Kopka J, Sung DY, Zhao W, Popp M, Porat R, Guy CL (2007) Transcript and metabolite profiling during cold acclimation of Arabidopsis reveals an intricate relationship of cold-regulated gene expression with modifications in metabolite content. Plant J 50: 967-981

Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S, Darzentas N, Kunin V, Lopez-Bigas N (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. Nucleic Acids Res 33: 6083-6089

Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gill M, Karp PD (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. Nucleic Acids Res 33: D334-337

Keurentjes JJB, Jingyuan F, de Vos CHR, Lommen A, Hall RD, Bino RJ, van der Plas LHW, Jansen RC, Vreugdenhil D, Koornneef M (2006) The genetics of plant metabolism. Nature Genetics 38: 842-849

Kim JK, Bamba T, Harada K, Fukusaki E, Kobayashi A (2007) Time-course metabolic profiling in *Arabidopsis thaliana* cell cultures after salt stress treatment. J Exp Bot 58: 415–424

Klotke J, Kopka J, Gatzke N, Heyer AG (2004) Impact of soluble sugar concentrations on the acquisition of freezing tolerance in accessions of Arabidopsis thaliana with contrasting cold adaptation. Plant Cell Environ 27:1395-1404

Kopka J, Fernie AF, Weckwerth W, Gibon Y, Stitt M (2004) Metabolite profiling in plant biology: Platforms and destinations. Genome Biol 5: 109-117

Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmüller E, Dörmann P, Gibon Y, Stitt M, Willmitzer L, Fernie AR, Steinhauser D (2005) GMD@CSBDB: The Golm Metabolome Database. Bioinformatics 21: 1635-1638

Kopka J (2006a) Gas chromatography mass spectrometry. In: Saito K, Dixon RA, Willmitzer L (eds) Plant metabolomics. Biotechnology in agriculture and forestry Vol. 57, Nagata T, Loerz H, Widholm JM (eds), Springer-Verlag, Berlin Heidelberg New York, pp 3-20

Kopka J (2006b) Current challenges and developments in GC-MS based metabolite profiling technology. J Biotechnol 124: 312-322

Knapp DR (1979) Handbook of analytical derivatization reactions. Wiley Publishers, New York

Kremsky J (2005) PubChem versus CAS. Chem Engineering News 83: 6

Kummel A, Panke S, Heinemann M (2006) Systematic assignment of thermodynamic constraints in metabolic network models. BMC Bioinformatics 7: 512

Lange, BM, Ghassemian, M (2005) Comprehensive post-genomic data analysis approaches integrating biochemical pathway maps. Phytochemistry 66: 413-451

Larkindale J, Hall JD, Knight MR, Vierling E (2005) Heat stress phenotypes of *Arabidopsis* mutants implicate multiple signaling pathways in the acquisition of thermotolerance. Plant Physiol 138: 882–888

Levitt J (1972) Responses of plants to environmental stresses. Academic Press, New York

Lindon JC, Keun HC, Ebbels TMD, Pearce JMT, Holmes E, Nicholson JK (2005) The Consortium for Metabonomic Toxicology (COMET): aims, activities and achievements. Pharmacogenomics 6: 691-699

Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie AR (2006) Gas chromatography mass spectrometry-based metabolite profiling in plants. Nat Protocols 1: 387-396

Lisso J, Steinhauser D, Altmann T, Kopka J, Muessig C (2005) Identification of brassinosteroid-related genes by means of transcript co-response analyses. Nucleic Acids Research 33: 2685-2696

Lohse S, Schliemann W, Ammer C, Kopka J, Strack D, Fester T (2005) Organization and metabolism of plastids and mitochondria in arbuscular mycorrhizal roots of Medicago truncatula. Plant Physiology 139: 329-340

Luedemann A, Weicht D, Selbig J, Kopka J (2004) Pathway visualization and editing system. Bioinformatics 20: 2841-2844

Luedemann A, Erban A, Wagner C, Kopka J (2005) Method for analyzing microbial metabolites by resolving isotopic mass differences within one metabolite using MALDI-TOF. PCT Int. Appl. (2005), 1050 pp. WO 2005059556 A1 20050630 CAN 143:74456 AN 2005:564800

Luedemann A, Strassburg K, Erban A, Kopka J (2008) TagFinder for the quantitative analysis of gas chromatography - mass spectrometry (GC-MS) based metabolite profiling experiments. Bioinformatics 24: 732 -737

Lu HM, Dunn WB, Shen HL, Kell DB, Liang YZ (2008) Comparative evaluation of software for deconvolution of metabolomics data based on GC-TOF-MS. Trends Anal Chem 27: 215-227

MacLeod JK, Flanigan IL, Williams JF, Collins JG (2001) Mass spectrometric studies of the path of carbon in photosynthesis: positional isotopic analysis of C-13-labeled C-4 to C-7 sugar phosphates. J Mass Spectrometry 36: 500-508

Marriott P, Shellie R, Fergeus J, Ong R, Morrison P (2000) High resolution essential oil analysis by using comprehensive gas chromatographic methodology. Flav Fragr J 15: 225-239

Marriott P, Shellie R (2002) Principles and applications of comprehensive two-dimensional gas chromatography. Trends Anal Chem 21: 573-583

Mashego MR, Wu L, Van Dam JC, Ras C, Vinke JL, Van Winden WA, Van Gulik WM, Heijnen JJ (2004) MIRACLE: mass isotopomer ratio analysis of U-C-13-labeled extracts. A new method for accurate quantification of changes in concentrations of intracellular metabolites. Biotechnol Bioeng 85: 620-628

Masoudi-Nejad A, Goto S, Jauregui R, Ito M, Kawashima S, Moriya Y, Endo TR, Kanehisa M (2007) EGENES: Transcriptome-based plant database of genes with metabolic pathway information and expressed sequence tag indices in KEGG. Plant Physiol 144: 857-866

Mehrotra B, Mendes P (2006) Bioinformatics: Approaches to integrate metabolomics and other systems biology data. In: Saito K, Dixon RA, Willmitzer L (eds) Plant metabolomics. Biotechnology in agriculture and forestry Vol. 57, Nagata T, Loerz H, Widholm JM (eds), Springer-Verlag, Berlin Heidelberg New York, pp 105-116

Meyer RC, Steinfath M, Lisec J, Becher M, Witucka-Wall H, Toerjek O, Fiehn O, Eckardt A, Willmitzer L, Selbig J, Altmann T (2007) The metabolic signature related to high plant growth rate in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 104: 4759-4764

Moco S, Bino RJ, Vorst O, Verhoeven HA, de Groot J, van Beek TA, Vervoort J, de Vos CHR (2006) A Liquid Chromatography Mass Spectrometry based Metabolome Database for Tomato. Plant Physiol 141: 1205-1218

Mueller A, Duechting P, Weiler EW (2002) A multiplex GC-MS/MS technique for the sensitive and quantitative single-run analysis of acidic phytohormones and related compounds, and its application to *Arabidopsis thaliana*. Planta 216:44-56

Munns R (2002) Comparative physiology of salt and water stress. Plant Cell Environ 25: 239-250

Munns R (2005) Genes and salt tolerance: bringing them together. New Phytol 167: 645-663

Nicholson JK, Lindon JC, Holmes E (1999) 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 29: 1181-1189

Nicholson JK, Connelly J, Lindon JC, Holmes E (2002) Metabonomics: a platform for studying drug toxicity and gene function. Nat Rev Drug Discov 1: 153-161

Nielsen J, Oliver S (2005) The next wave in metabolome analysis. Trends Biotechnol 23: 544–546

Nobeli I, Krissinel EB, Thornton JMB (2003) A structure-based anatomy of the *E. coli* metabolome. J Mol Biol 334: 697–719

O'Hagan S, Dunn WB, Brown M, Knowles JD, Kell DB (2005) Closed-loop, multiobjective optimization of analytical instrumentation: Gas chromatography/time-of-flight mass spectrometry of the metabolomes of human serum and of yeast fermentations. Anal Chem 77: 290-303

Oliver SG, Winson MK, Kell DB, Baganz F (1998) Systematic functional analysis of the yeast genome. Trends Biotechnol 16: 373-378

Panikulangara TJ, Eggers-Schumacher G, Wunderlich M, Stransky H, Schöffl F (2004) Galactinol synthase1. A novel heat shock factor target gene responsible for heat-induced synthesis of raffinose family oligosaccharides in *Arabidopsis*. Plant Physiol 136: 3148–3158

Pinheiro C, Passarinho JA, Pinto Ricardo C (2004) Effect of drought and rewatering on the metabolism of *Lupinus albus* organs. J Plant Physiol 161: 1203–1210

Pool WG, de Leeuw JW, van de Graaf B (1996) Backfolding applied to differential gas chromatography/mass spectrometry as a mathematical enhancement of chromatographic resolution. J Mass Spectrom 31: 509-516

Pool WG, de Leeuw JW, van de Graaf B (1997a) Automated extraction of pure mass spectra from gas chromatographic/mass spectrometric data. J Mass Spectrom 32: 438-443

Pool WG, de Leeuw JW, van de Graaf B (1997b) Automated processing of GC/MS data: quantification of the signals of individual components. J Mass Spectrom 32: 1253-1257

Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff, HV, van Dam K, Oliver SG (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nat Biotechnol 19: 45-50

Ratcliffe RG, Shachar-Hill Y (2006) Measuring multiple fluxes through plant metabolic networks. Plant J 45: 490-511

Rautengarten C, Steinhauser D, Bussis D, Stintzi A, Schaller A, Kopka J, Altmann T (2005) Inferring hypotheses on functional relationships of genes: Analysis of the Arabidopsis thaliana subtilase gene family. PLOS Computational Biology 1: 297-312

Rizhsky L, Liang HJ, Shuman J, Shulaev V, Davletova S, Mittler R (2004) When defence pathways collide. The response of Arabidopsis to a combination of drought and heat stress. Plant Physiol 134: 1683–1696

Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L (2000) Technical advance: Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. Plant J 23: 131-142

Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L, Fernie A (2001a) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. Plant Cell 13: 11-29

Roessner U, Willmitzer L, Fernie AR (2001b) High-resolution metabolic phenotyping of genetically and environmentally diverse plant systems - identification of phenocopies. Plant Physiol 127: 749-764

Roessner U, Willmitzer L, Fernie AR (2002) Metabolic profiling and biochemical phenotyping of plant systems. Plant Cell Rep 21: 189-196

Ryan D, Shellie R, Tranchida P, Casilli A, Mondello L, Marriott P (2004) Analysis of roasted coffee bean volatiles by using comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry. J Chromatogr A 1054: 57-65

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J (2003) TM4: A free, open-source system for microarray data management and analysis. Biotechniques 34: 374-378

Saeed AI, Hagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li JW, Thiagarajan M, White JA, Quackenbush J (2006) TM4 microarray software suite. Methods Enzymol 411: 134-193

Sanchez DH, Lippold F, Redestig H, Hannah M, Erban A, Kraemer U, Kopka J, Udvardi MK (2008a) Integrative functional genomics of salt acclimation in the model legume *Lotus japonicus*. Plant J 53: 973-987

Sanchez DH, Siahpoosh MR, Roessner U, Udvardi MK, Kopka J (2008b) Plant metabolomics reveals conserved and divergent metabolic responses to salinity. Physiologica Plantarum 132: 209-219

Sanchez DH, Redestig H, Kraemer U, Udvardi MK, Kopka J (2008c) Metabolome-ionome-biomass interactions: What can we learn about salt stress by multiparallel phenotyping? Plant Signaling and Behavior 3: 1-3

Sauter H, Lauer M, Fritsch H (1988) Metabolite profiling of plants - a new diagnostic technique. Abstr Pap Am Chem Soc 195: 129

Schaarschmidt S, Kopka J, Ludwig-Mueller J, Hause B (2007) Regulation of arbuscular mycorrhization by apoplastic invertases: enhanced invertase activity in the leaf apoplast affects the symbiotic interaction. Plant Journal 51: 390-405

Schaefer J, Stejskal EO, Beard CF (1975) C-13 Nuclear magnetic resonance analysis of metabolism in soybeans labeled by $^{13}CO_2$. Plant Physol 55: 1048-1053

Schaefer J, Kier LD, Stejskal EO (1980) Characterization of photorespiration in intact leaves using C-13 dioxide labeling. Plant Physiol 65: 254-259

Schauer N, Steinhauser D, Strelkov S, Schomburg D, Allison G, Moritz T, Lundgren K, Roessner-Tunali U, Forbes MG, Willmitzer L, Fernie AR, Kopka J (2005) GC-MS libraries for the rapid identification of metabolites in complex biological samples. FEBS Lett 579: 1332-1337

Schauer N, Semel Y, Roessner U, Gur A, Balbo I, Carrari F, Pleban T, Perez-Melis A, Bruedigam C, Kopka J, Willmitzer L, Zamir D, Fernie AR (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. Nat Biotechnol 24: 447-454

Schwall K, Zielenbach K (2000) SciFinder - A new generation of research tool. Chem Innovat 30: 45-50

Schwender J, Goffman F, Ohlrogge JB, Shachar-Hill Y (2004) Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds. Nature 432: 779-782

Schwender J, Ohlrogge J, Shachar-Hill Y (2004) Understanding flux in plant metabolic networks. Curr Opin Plant Biol 7: 309–317

Shang S, Tan DS (2005) Advancing chemistry and biology through diversity-oriented synthesis of natural product-like libraries. Curr Opin Chem Biol 9: 248-258

Shao XG, Wang GQ, Wang SF, Su QD (2004) Extraction of mass spectra and chromatographic profiles from overlapping GC/MS signal with background. Anal Chem 76: 5143-5148

Shastri AA, Morgan JA (2007) A transient isotopic labeling methodology for ^{13}C metabolic flux analysis of photoautotrophic microorganisms. Phytochem 68: 2302-2312

Shellie R, Marriott P, Morrison P (2001) Concepts and preliminary observations on the triple-dimensional analysis of complex volatile samples by using GCxGC-TOFMS. Anal Chem 73: 1336-1344

Sinha AE, Fraga CG, Prazen BJ, Synovec RE (2004a) Trilinear chemometric analysis of twodimensional comprehensive gas chromatography-time-of-flight mass spectrometry data. J Chromatogr A 1027: 269-277

Sinha AE, Hope JL, Prazen BJ, Nilsson EJ, Jack RM, Synovec RE (2004b) Algorithm for locating analytes of interest based on mass spectral similarity in GC × GC–TOF-MS data: analysis of metabolites in human infant urine. J Chromatogr A 1058: 209-215

Sinha AE, Prazen BJ, Synovec RE (2004c) Trends in chemometric analysis of comprehensive two-dimensional separations. Anal Bioanal Chem 378:1948-1951

Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G (2006) XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem 78: 779-787

Spasić I, Dunn WB, Velarde G, Tseng A, Jenkins H, Hardy NW, Oliver SG, Kell DB (2006) MeMo: a hybrid SQL/XML approach to metabolomic data management for functional genomics. BMC Bioinformatics 7: 281

Stein SE (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/ mass spectrometry data. J Am Soc Mass Spectrom 10: 770-781

Steinhauser D, Junker BH, Luedemann A, Selbig J, Kopka J (2004a) A hypothesis-driven approach to predict transcriptional units from gene expression data. Bioinformatics 20: 1928-1939

Steinhauser D, Usadel B, Luedemann L, Thimm O, Kopka J (2004b) CSB.DB: A comprehensive systems-biology database. Bioinformatics 20: 3647-3651

Steinhauser D, Kopka J (2007) Methods, applications and concepts of metabolite profiling: primary metabolism. In: Fernie AR, Baginsky S (eds) Plant systems biology. Experimentia Supplementum Vol. 97, Verlag Birkhäuser, Basel, pp 171-194

Stephanopoulos G, Alper H, Moxley J (2004) Exploiting biological complexity for strain improvement through systems biology. Nat Biotechnol 22: 1261-1267

Steuer R, Kurths J, Fiehn O, Weckwerth W (2003) Observing and interpreting correlations in metabolomic networks. Bioinformatics 19: 1019-1026

Strehmel N, Hummel J, Erban A, Strassburg K, Kopka J (2008) Estimation of retention index thresholds for compound matching using routine gas chromatography-mass spectrometry based metabolite profiling experiments. J Chromatogr B 871: 182-190

Strelkov S, von Elstermann M, Schomburg D (2004) Comprehensive analysis of metabolites in *Corynebacterium glutamicum* by gas chromatography/mass spectrometry. Biol Chem 385: 853-861

Sumner LW, Mendes P, Dixon RA (2003) Plant metabolomics: Large-scale phytochemistry in the functional genomics era. Phytochemistry 62: 817-836

Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, Fan TWM, Fiehn O, Goodacre R, Griffin JL, Hankemeier T, Hardy N, Harnly J, Higashi R, Kopka J, Lane AN, Lindon JC, Marriott P, Nicholls AW, Reily MD, Thaden JJ, Viant MR (2007) Proposed minimum reporting standards for chemical analysis. Metabolomics 3: 211-221

Thomashow MF (1999) Plant cold acclimation: freezing tolerance genes and regulatory mechanisms. Annu Rev Plant Physiol Plant Mol Biol 50: 571–599

Tikunov YM, Lommen A, de Vos CH, Verhoeven HA, Bino RJ, Hall RD, Bovy AG (2005) A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. Plant Physiol 139: 1125-1137

Tikunov YM, Verstappen FWA, Hall RD (2007) Metabolomic profiling of natural volatiles: Headspace trapping: GC-MS. In: Weckwerth W (ed) Metabolomics: methods and protocols. Humana Press, Totowa, pp 19-38

Trethewey RN, Krotzky AJ, Willmitzer L (1999) Metabolic profiling: a Rosetta stone for genomics? Curr Opin Plant Biol 2: 83-85

Trethewey RN (2004) Metabolite profiling as an aid to metabolic engineering in plants. Curr Opin Plant Biol 7: 196-201

Trygg J, Wold S (2002) Orthogonal projections to latent structures (O-PLS). J Chemom 16: 119-28

Tweeddale H, Notley-McRobb L, Ferenci T (1998) Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool ("metabolome") analysis. J Bacteriol 180: 5109-5116

Urbanczyk-Wochniak E, Luedemann A, Kopka J, Selbig J, Roessner-Tunali U, Willmitzer L, Fernie AR (2003) Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. EMBO Reports 4: 989-993

Usadel B, Nagel A, Thimm O, Redestig H, Blaesing OE, Palacios-Rojas N, Selbig J, Hannemann J, Piques MC, Steinhauser D, Scheible WR, Gibon Y, Morcuende R, Weicht D, Meyer S, Stitt M (2005) Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of corresponding genes, and comparison with known responses. Plant Physiol 138: 1195-1204

van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics 7: 142

van den Dool H, Kratz PD (1963) A generalization of the retention index system including linear temperature programmed gas-liquid partition chromatography. J Chromatogr 11: 463-471

van Deursen MM, Beens J, Janssen HG, Leclercq PA, Cramers CA (2000) Evaluation of time-of-flight mass spectrometric detection for fast gas chromatography. J Chromatogr A 878: 205-213

Veriotti T, Sacks R (2000) High speed GC/MS of gasoline-range hydrocarbon compounds using a pressure-tunable column ensemble and time-of-flight detection. Anal Chem 72: 3063-3069

Veriotti T, Sacks R (2001) High-speed GC and GC/time-of-flight MS of lemon and lime oil samples. Anal Chem 73: 4395-4402

Vorst O, de Vos CHR, Lommen A, Staps RV, Visser RGF, Bino RJ, Hall RD (2005) A non-directed approach to the differential analysis of multiple LCMS derived metabolic profiles. Metabolomics 1: 169–180

Vreuls RJJ, Dalluge J, Brinkman UAT (1999) Gas chromatography time-of-flight mass spectrometry for sensitive determination of organic microcontaminants. J Microcolumn Sep 11: 663-675

Wagner C, Sefkow M, Kopka J (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. Phytochem 62: 887-900

Weckwerth W, Wenzel K, Fiehn O (2004a) Process for the integrated extraction identification, and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. Proteomics 1: 78-83

Weckwerth W, Loureiro ME, Wenzel K, Fiehn O (2004b) Differential metabolic networks unravel the effects of silent plant phenotypes. Proc Natl Acad Sci USA 101: 7809-7814

Whitley KM (2002) Analysis of SciFinder scholar and web of science citation searches. J Amer Soc Informat Sci Technol 53: 1210-1215

Wishart DS (2007) Human Metabolome Database: completing the 'human parts list'. Pharmacogenomics 8 (7): 683-686

Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, Fung C, Nikolai L, Lewis M, Coutouly MA, Forsythe I, Tang P, Shrivastava S, Jeroncic K, Stothard P, Amegbey G, Block D, Hau DD, Wagner J, Miniaci J, Clements M, Gebremedhin M, Guo N, Zhang Y, Duggan GE, MacInnis GD, Weljie AM, Dowlatabadi R, Bamforth F, Clive D, Greiner R, Li L, Marrie T, Sykes BD, Vogel HJ, Querengesser L (2007) HMDB: The human metabolome database. Nucleic Acids Res 35: D521-D526

Wu L, Mashego MR, van Dam JC, Proell AM, Vinke JL, Ras C, van Winden WA, van Gulik WM, Heijnen JJ (2005) Quantitative analysis of the microbial metabolome by isotope dilution mass spectrometry using uniformly C-13-labeled cell extracts as internal standards. Anal Biochem 336: 164-171

Wurtele ES, Li J, Diao LX, Zhang HL, Foster CM, Fatland B, Dickerson U, Brown A, Cox Z, Cook D, Lee EK, Hofmann H (2003) MetNet: Software to Build and Model the Biogenetic Lattice of Arabidopsis. Compar Funct Genom 4: 239-245

Zhang PF, Foerster H, Tissier CP, Mueller L, Paley S, Karp PD, Rhee SY (2005) MetaCyc and AraCyc. Metabolic pathway databases for plant research. Plant Physiol 138: 27-37

Zuther E, Büchel K, Hundertmark M, Stitt M, Hincha DK, Heyer AG (2004) The role of raffinose in the cold acclimation response of *Arabidopsis thaliana*. FEBS Lett 576: 169–173

Zuther E, Koehl K, Kopka J (2007) Comparative metabolome analysis of the salt response in breeding cultivars of rice. In: Jenks MA, Hasegawa PM, Jain SM (eds) Advances in molecular breeding toward drought and salt tolerant crops. Springer-Verlag, Berlin, Heidelberg, New York, pp 285-315

ABSTRACT

The uptake of nutrients and their subsequent chemical conversion by reactions which provide energy and building blocks for growth and propagation is a fundamental property of life. This property is termed metabolism. In the course of evolution life has been dependent on chemical reactions which generate molecules that are common and indispensable to all life forms. These molecules are the so-called primary metabolites. In addition, life has evolved highly diverse biochemical reactions. These reactions allow organisms to produce unique molecules, the so-called secondary metabolites, which provide a competitive advantage for survival. The sum of all metabolites produced by the complex network of reactions within an organism has since 1998 been called the metabolome. The size of the metabolome can only be estimated and may range from less than 1,000 metabolites in unicellular organisms to approximately 200,000 in the whole plant kingdom. In current biology, three additional types of molecules are thought to be important to the understanding of the phenomena of life: (1) the proteins, in other words the proteome, including enzymes which perform the metabolic reactions, (2) the ribonucleic acids (RNAs) which constitute the so-called transcriptome, and (3) all genes of the genome which are encoded within the double strands of desoxyribonucleic acid (DNA). Investigations of each of these molecular levels of life require analytical technologies which should best enable the comprehensive analysis of all proteins, RNAs, et cetera. At the beginning of this thesis such analytical technologies were available for DNA, RNA and proteins, but not for metabolites. Therefore, this thesis was dedicated to the implementation of the gas chromatography – mass spectrometry technology, in short GC-MS, for the in-parallel analysis of as many metabolites as possible. Today GC-MS is one of the most widely applied technologies and indispensable for the efficient profiling of primary metabolites. Depending on the biological sample, a single GC-MS profile typically covers 50 – 150 known metabolites and allows the facultative discovery of hundreds of unexpected or yet non-identified metabolites. Besides such qualitative assessments of metabolite composition, the relative changes of metabolite concentrations can be characterized by GC-MS based profiling. The GC-MS based profiling method has been continuously improved during my thesis with the ultimate aim to advance our understanding of gene function and metabolic responses in higher plants.

The main achievements and research topics of this work can be divided into technological advances and novel insights into the metabolic mechanisms which allow plants to cope with environmental stresses. Firstly, the GC-MS profiling technology has been highly automated and standardized both, at the level of laboratory chemistry and at the level of data processing. The major technological achievements were (1) substantial contributions to the development of automated and, within the limits of GC-MS, comprehensive chemical analysis, (2) contributions to the implementation of time of flight mass spectrometry for GC-MS based metabolite profiling, (3) the creation of a software platform for reproducible GC-MS data processing, named TagFinder, and (4) the establishment of an internationally coordinated library of mass spectra which allows the identification of metabolites in diverse and complex biological samples. In addition, the Golm Metabolome Database (GMD) has been initiated to harbor this library and to cope with the increasing amount of generated profiling data. This database makes publicly available all chemical information essential for GC-MS profiling and has been extended towards a global resource of GC-MS based metabolite profiles. Querying the concentration changes of hundreds of known and yet non-identified metabolites has recently been enabled by uploading standardized, TagFinder-processed data. Long-term technological aims have been pursued with the central aims (1) to enhance the precision of absolute and relative quantification and (2) to enable the combined analysis of metabolite concentrations and metabolic flux. In contrast to concentrations which provide information on metabolite amounts, flux analysis provides information on the speed of biochemical reactions or reaction sequences, for example on the rate of CO2 conversion into metabolites. This conversion is an essential function of plants which is the basis of life on earth. Secondly, GC-MS based metabolite profiling technology has been continuously applied to advance plant stress physiology. These efforts have yielded a detailed description of and new functional insights into metabolic changes in response to high and low temperatures as well as common and divergent responses to salt stress among higher plants, such as Arabidopsis thaliana, Lotus japonicus and rice (Oryza sativa). Time course analysis after temperature stress and investigations into salt dosage responses indicated that metabolism changed in a gradual manner rather than by stepwise transitions between fixed states. In agreement with these observations, metabolite profiles of the model plant Lotus japonicus, when exposed to increased soil salinity, were demonstrated to have a highly predictive power for both NaCl accumulation and plant biomass. Thus, it may be possible in the future to use GC-MS based metabolite profiling as a breeding tool to support the selection of individual plants that cope best with salt stress or other environmental challenges.

ZUSAMMENFASSUNG

Die Aufnahme von Nährstoffen und ihre chemische Umwandlung mittels Reaktionen, die Energie und Baustoffe für Wachstum und Vermehrung bereitstellen, ist eine grundlegende Eigenschaft des Lebens. Diese Eigenschaft wird Stoffwechsel oder, wie im Folgenden, Metabolismus genannt. Im Verlauf der Evolution war alles Leben abhängig von solchen Reaktionen, die essentielle und allen Lebensformen gemeinsame Moleküle erzeugen. Über diese sogenannten Primärmetabolite hinaus sind hochdiverse Reaktionen entstanden. Diese erlauben Organismen, einzigartige sogenannte Sekundärmetabolite zu produzieren, die in der Regel einen zusätzlichen Überlebensvorteil vermitteln. Die Gesamtheit aller Metabolite, die von dem komplexen Reaktionsnetzwerk in Organismen erzeugt werden, nennt man seit 1998 das Metabolom. Die Größe des Metaboloms kann nur geschätzt werden und variiert von weniger als 1.000 Metabolite in einzelligen Organismen bis zu ~200.000 im gesamten Pflanzenreich. Neben der Gesamtheit aller Metabolite werden heute drei weitere Arten an Molekülen als wesentlich betrachtet, um die Phänomene des Lebens zu verstehen: erstens die Proteine, deren Summe, das Proteom, auch die Enzyme einschließt, die die obigen metabolischen Reaktionen durchführen, zweitens die Ribonukleinsäuren (RNS), deren Gesamtheit als Transkriptom bezeichnet wird, und drittens die doppelsträngige Desoxyribonukleinsäure (DNS), die das Genom, die Summe aller Gene eines Organismus, ausmacht. Die Untersuchung aller dieser vier molekularen Ebenen des Lebens erfordert Technologien, die idealerweise die vollständige Analyse der Gesamtheit aller DNS-, RNS-, Protein-Moleküle, bzw. Metabolite erlauben. Zu Beginn meiner Arbeiten waren solche Technologien für DNS, RNS, und Proteine verfügbar, aber nicht für Metabolite. Aus diesem Grund habe ich meine Forschungstätigkeit auf das Ziel ausgerichtet, so viele Metabolite wie irgend möglich in einer gemeinsamen Analyse zu erfassen. Zu diesem Zweck habe ich mich auf eine einzelne Technik, nämlich die gekoppelte Gaschromatographie und Massenspektrometrie, kurz GC-MS, konzentriert. Nicht zuletzt durch meine Arbeiten ist GC-MS heute eine der am häufigsten angewandten Technologien und unverzichtbar für das breite Durchmustern der Metabolite. In Abhängigkeit von der Wahl der biologischen Probe deckt eine einzige solche GC-MS-Profilanalyse 50-150 bekannte Primärmetabolite ab. Über diese hinaus können mit der gleichen GC-MS-Analyse noch hunderte von unerwarteten oder sogar noch nicht identifizierten Metabolite entdeckt werden. Neben der qualitativen Bestandsaufnahme der Metabolitzusammensetzung einer Probe können zudem die Veränderungen in der Menge jedes einzelnen beobachteten Metabolits erfasst werden. Nach der Etablierung der grundlegenden GC-MS-Profilanalyse-Technologie habe ich diese im Verlauf meiner Arbeiten kontinuierlich erweitert und verbessert. Das angewandte Ziel dieser Arbeiten war und ist es, das Wissen über das Zusammenwirken von Genfunktion und Metabolismus am Beispiel physiologischer Reaktionen auf Umweltstress zu vertiefen.

Die Haupterrungenschaften meiner Arbeiten liegen sowohl in den technischen Neuerungen als auch in den Einsichten in metabolische Mechanismen, die es Pflanzen erlauben, erfolgreich auf Umwelteinflüsse zu reagieren. Die technologischen Errungenschaften waren erstens wesentliche Beiträge zur Labor-Automatisierung und zur Auswertung von modernen, auf Flugzeitmassenspektrometrie beruhenden, GC-MS-Profilanalysen, zweitens die Entwicklung einer entsprechenden Prozessierungs-Software, genannt TagFinder, und drittens die Etablierung einer internationalen Datensammlung zur Metabolitidentifizierung aus komplexen Mischungen. Diese massenspektralen und gaschromatographischen Daten haben seit 2005 Eingang in die von mir initiierte Entwicklung der Golm Metabolom Datenbank (GMD) gefunden, die die zunehmend wachsenden GC-MS-Referenzdaten wie auch die Metabolitprofildaten verwaltet und öffentlich zugänglich macht. Darüber hinaus wurden die langfristigen Ziele einer verbesserten Präzision für relative und absolute Quantifizierung wie auch einer Kopplung von Konzentrationsbestimmung und metabolischen Flussanalysen mittels GC-MS verfolgt. Sowohl die Stoffmengen als auch die Geschwindigkeit der Stoffaufnahme und der chemischen Umsetzung, d.h. der metabolische Fluss, sind wesentlich für neue biologische Einsichten. In diesem Zusammenhang wurde von mir die Aufnahme von CO2 durch Pflanzen, der Basis allen Lebens auf der Erde, untersucht. Angewandt auf das Temperaturstress- und Salzstressverhalten von Modell- und Kulturpflanzen, nämlich des Ackerschmalwands (Arabidopsis thaliana), des Hornklees (Lotus japonicus) und der global bedeutendsten Nutzpflanze Reis (Oryza sativa), wurden detaillierte und vergleichende neue metabolische Einsichten in den Zeitverlauf der Temperaturanpassung und die Anpassung an zunehmend salzhaltige Böden erzielt. Metabolismus verändert sich unter diesen Bedingungen allmählich fortschreitend und nicht in plötzlichen Übergängen. Am Beispiel des Hornklees konnte gezeigt werden, dass Metabolitprofilanalysen eine hohe Vorhersagekraft für die Biomasseerzeugung unter Salzeinfluss wie auch für die Aufnahme von Salz durch die Pflanze haben. So mag es in Zukunft möglich werden, GC-MS-Profilanaysen anzuwenden, um den Züchtungsprozess von Kulturpflanzen zu beschleunigen und die Auswahl der bestgeeigneten Nachkommenschaft aus neuen Kreuzungen zu unterstützen.

CURRICULUM VITAE

Name	Dr. rer. nat. **Joachim Kopka**
Date of birth	4th May 1962, Münster
Nationality	German
Resident	Stubbenkammer Str. 2, D-10437 Berlin, Germany
Date of marriage	26th August 1992, Münster
	to Andrea Kopka, né Scholthaus
Children	Johanna Kopka, 5th September 1995
	Jonas Kopka, 3rd May 1997

Education and career

1968 - 1972	Thomas-Morus primary school, Münster
1972 - 1981	Pascal-Gymnasium, Münster
1981 - 1982	Public service at the medical hospital of Münster-Hiltrup
1982 - 1988	**Study of biology and chemistry**, Westfälische Wilhems-Universität (WWU), Münster, and state exam Sekundarstufen I and II of biological and chemical sciences with the thesis “The substantial composition of pulmonate eggs”, Prof. W. Bottke, Department of Zoology, WWU, Münster
1989 - 1992	**PhD thesis** (magna cum laude) "Characterisation of acyl carrier proteins from *Cuphea*-plants and their role in the synthesis of middle chain fatty acids", Prof. F. Spener, Department of Biochemistry, WWU, Münster
1992 - 1993	**Postdoctoral research assistant**, “Analysis of fatty acid de-novo-synthesis in soybean, *Glycine max* (L.) Merr.", Prof. J.G. Jaworski, Department of Chemistry, Miami-University, Oxford, OH, USA
1994 - 1996	**Postdoctoral research assistant**, “Molecular biology of the phosphoinositide signal transduction in potato plants, *Solanum tuberosum* cv. Désirée”, Dr. B. Müller-Röber, Institut für Genbiologische Forschung (IGF), Berlin, Germany
1996 - 1997	**Postdoctoral research assistant**, "Production of polyhydroxy fatty acids targeted to the mitochondria in crops", Dr. R.N. Trethewey, Max-Planck-Institute of Molecular Plant Physiology (MPIMP), Potsdam-Golm, Germany
1997 - 1998	**Postdoctoral research assistant**, “Development of a GC/MS technology for the metabolic screening of plant samples", Dr. R.N. Trethewey, MPIMP, Potsdam-Golm

1998 - 2000	Leave from the MPIMP to co-found the Metanomics GmbH & Co. KGaA company (Metanomics), Berlin
1998 - 1999	**Leader of the technical center** Bioanalytics, Metanomics, Berlin
1999 - 2000	**Leader of the technical center** Bioinformatics/ Datamining, Metanomics, Berlin
2001 -	**Research group leader**, Department of Molecular Plant Physiology, Prof. L. Willmitzer, MPIMP, Potsdam-Golm
2002 - 2008	**Employees´ representative** of the MPIMP at the section of biological and medical sciences of the Max-Planck-Society
2004 - 2007	**Member of the advisory committee**, Bioinformatics Center Gatersleben-Halle (BIC-GH)
2005	Advanced training course according to § 15 Abs. 2 Nr. 3 (Gentechnik-Sicherheitsverordnung)
2005 -	**Monitoring editor** of the "Plant Physiology" journal
2007	**Guest editor** of the "Physiologia Plantarum" special issue, volume 132, on metabolomics and metabolism
2008 -	**Member of the DECHEMA e.V.** (Gesellschaft für Chemische Technik und Biotechnologie e.V., Mitgl. Nr. 43807)

Publications

2008

Böttcher C, Centeno D, Freitag J, Hoefgen R, Koehl K, **Kopka J**, Kroymann J, Matros A, Mock H-P, Neumann S, Pfalz M, von Roepenack-Lahaye E, Schauer N, Trenkamp S, Zubriggen M, Fernie AR (2008) Teaching (and learning from) metabolomics: The 2006 PlantMetaNet ETNA metabolomics research school. *Physiologica Plantarum* 132 (2): 136-149

Cartieaux F, Contesto C, Gallou A, Desbrosses G, **Kopka J**, Taconnat L, Renou JP, Touraine B (2008) Simultaneous interaction of *Arabidopsis thaliana* with *Bradyrhizobium sp*. strain ORS278 and *Pseudomonas syringae* pv. tomato DC3000 leads to complex transcriptome changes. *Molecular Plant Microbe Interactions* 21(2): 244-259

Eisenhut M, Huege J, Schwarz D, Bauwe H, **Kopka J**, Hagemann M (2008) Metabolome phenotyping of inorganic carbon limitation in cells of the wild type and photorespiratory mutants of the cyanobacterium *Synechocystis sp.* strain PCC 6803. *Plant Physiology* (First published on October 22, 2008; doi:10.1104/pp.108.129403)

Guy CL, Kaplan F, **Kopka J**, Selbig J, Hincha D (2008) Metabolomics of temperature stress. *Physiologica Plantarum* 132 (2): 220-235

Hoehenwarter W, van Dongen JT, Wienkoop S, Steinfath M, Hummel J, Erban A, Sulpice R, Regierer B, **Kopka J**, Geigenberger P, Weckwerth W (2008) A rapid approach for phenotype-screening and database

independent detection of cSNP/protein polymorphism using mass accuracy precursor alignment. *Proteomics* 8(20): 4214-4225

Kempa S, Krasensky J, Dal Santo S, **Kopka J**, Jonak C (2008) A central role of abscisic acid in stress-regulated carbohydrate metabolism. *PloS ONE* 3(12): e3935 (doi:10.1371/journal.pone.0003935)

Leplé JC, Dauwe R, Morreel K, Storme V, Lapierre C, Pollet B, Naumann A, Kang KY, Kim H, Ruel K, Lefèbvre A, Joseleau JP, Grima-Pettenati J, De Rycke R, Andersson-Gunnerås S, Erban A, Fehrle I, Petit-Conil M, **Kopka J**, Polle A, Messens E, Sundberg B, Mansfield SD, Ralph J, Pilate G, Boerjan W (2008) Downregulation of cinnamoyl-coenzyme A reductase in poplar: Multiple-level phenotyping reveals effects on cell wall polymer metabolism and structure. *Plant Cell* 19 (11): 3669-3691

Levine LH, Kasahara H, Kopka J, Erban A, Fehrle I, Kaplan F, Zhao W, Littell RC, Guy C, Wheeler R, Sager J, Mills A, Levine HG (2008) Physiologic and metabolic responses of wheat seedlings to elevated and super-elevated carbon dioxide. *Advances in Space Research* 42: 1917–1928

Luedemann A, Strassburg K, Erban A, **Kopka J** (2008) TagFinder for the quantitative analysis of gas chromatography - mass spectrometry (GC-MS) based metabolite profiling experiments. *Bioinformatics* 24 (5): 732 -737

Sanchez DH, Lippold F, Redestig H, Hannah M, Erban A, Kraemer U, **Kopka J**, Udvardi MK (2008) Integrative functional genomics of salt acclimation in the model legume *Lotus japonicus*. *Plant Journal* 53 (6): 973-987

Sanchez DH, Redestig H, Kraemer U, Udvardi MK, **Kopka J** (2008) Metabolome-ionome-biomass interactions: What can we learn about salt stress by multiparallel phenotyping? Plant Signaling and Behavior 3 (8): 1-3

Sanchez DH, Siahpoosh MR, Roessner U, Udvardi MK, **Kopka J** (2008) Plant metabolomics reveals conserved and divergent metabolic responses to salinity. *Physiologica Plantarum* 132 (2): 209-219

Strehmel N, Hummel J, Erban A, Strassburg K, **Kopka J** (2008) Estimation of retention index thresholds for compound matching using routine gas chromatography-mass spectrometry based metabolite profiling experiments. *Journal of Chromatography B* 871: 182-190

Van Dongen JT, Fröhlich A, Ramírez-Aguilar SJ, Schauer N, Fernie AR, Erban A, **Kopka J**, Clark J, Langer A, Geigenberger P (2008) Transcript and metabolite profiling of the adaptive response to mild decreases in oxygen concentration in the roots of *Arabidopsis* plants. *Annals of Botany* (First published on August 6, 2008; doi:10.1093/aob/mcn126)

2007

Dauwe R, Morreel K, Goeminne G, Gielen B, Rohde A, Van Beeumen J, Ralph J, Boudet AM, **Kopka J**, Rochange SF, Halpin C, Messens E, Boerjan W (2007) Molecular phenotyping of lignin-modified tobacco reveals associated changes in cell-wall metabolism, primary metabolism, stress metabolism and photorespiration. *Plant Journal* 52 (2): 263-285

Hannah MA, Kraemer KM, Geffroy V, **Kopka J**, Blair MW, Erban A, Vallejos CE, Heyer AG, Sanders FET, Millner PA, Pilbeam DJ (2007) Hybrid weakness controlled by the dosage-dependent lethal (DL) gene

system in common bean (*Phaseolus vulgaris*) is caused by a shoot-derived inhibitory signal leading to salicylic acid-associated root death. *New Phytologist* 176 (3): 537-549

Hernandez G, Ramirez M, Valdes-Lopez O, Tesfaye M, Graham MA, Czechowski T, Schlereth A, Wandrey M, Erban A, Cheung F, Wu HC, Lara M, Town CD, **Kopka J**, Udvardi MK, Vance CP (2007) Phosphorus stress in common bean: Root transcript and metabolic responses. *Plant Physiology* 144 (2): 752-767

Huege J, Sulpice R, Gibon Y, Lisec J, Koehl K, **Kopka J** (2007) GC-EI-TOF-MS analysis of *in vivo*-carbon-partitioning into soluble metabolite pools of higher plants by monitoring isotope dilution after ($^{13}CO_2$)-labelling. *Phytochemistry* 68 (16-18): 2258-2272

Kaplan F, **Kopka J**, Sung DY, Zhao W, Popp M, Porat R, Guy CL (2007) Transcript and metabolite profiling during cold acclimation of Arabidopsis reveals an intricate relationship of cold-regulated gene expression with modifications in metabolite content. *Plant Journal* 50 (6): 967-981

Kempa S, Rozhon W, Samaj J, Erban A, Baluska F, Becker T, Haselmayer J, Schleiff E, **Kopka J**, Hirt H, Jonak C (2007) A plastid-localized glycogen synthase kinase 3 modulates stress tolerance and carbohydrate metabolism. *Plant Journal* 49 (6): 1076-1090

Parveen I, Moorby JM, Fraser MD, Allison GG, Kopka J (2007) Application of gas chromatography-mass spectrometry metabolite profiling techniques to the analysis of heathland plant diets of sheep. *Journal of Agricultural and Food Chemistry* 55 (4): 1129-1138

Schaarschmidt S, **Kopka J**, Ludwig-Mueller J, Hause B (2007) Regulation of arbuscular mycorrhization by apoplastic invertases: enhanced invertase activity in the leaf apoplast affects the symbiotic interaction. *Plant Journal* 51 (3): 390-405

Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, Fan TWM, Fiehn O, Goodacre R, Griffin JL, Hankemeier T, Hardy N, Harnly J, Higashi R, **Kopka J**, Lane AN, Lindon JC, Marriott P, Nicholls AW, Reily MD, Thaden JJ, Viant MR (2007) Proposed minimum reporting standards for chemical analysis. *Metabolomics* 3 (3): 211-221

2006

Domoney C, Duc G, Ellis THN, Ferrandiz C, Firnhaber C, Gallardo K, Hofer J, **Kopka J**, Kuster H, Madueno F, Munier-Jolain NG, Mayer K, Thompson R, Udvardi MK, Salon C (2006) Genetic and genomic analysis of legume flowers and seeds. *Current Opinion in Plant Biology* 9 (2): 133-141

Engelsberger WR, Erban A, **Kopka J**, Schulze WX (2006) Metabolic labeling of plant cell cultures with $K^{15}NO_3$ as a tool for quantitative analysis of proteins and metabolites. *Plant Methods* 2:14

Lisec J, Schauer N, **Kopka J**, Willmitzer L, Fernie AR (2006) Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nature Protocols* 1: 387 - 396

Niehl A, Lacomme C, Erban A, **Kopka J**, Kraemer U, Fisahn J (2006) Systemic Potato virus X infection induces defence gene expression and accumulation of beta-phenylethylamine-alkaloids in potato. *Functional Plant Biology* 33 (6): 593-604

Schauer N, Semel Y, Roessner U, Gur A, Balbo I, Carrari F, Pleban T, Perez-Melis A, Bruedigam C, **Kopka J**, Willmitzer L, Zamir D, Fernie AR (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nature Biotechnology* 24 (4): 447-454

Skirycz A, Reichelt M, Burow M, Birkemeyer C, Rolcik J, **Kopka J**, Zanor MI, Gershenzon J, Strnad M, Szopa J, Mueller-Roeber B, Witt I (2006) DOF transcription factor AtDof1.1 (OBP2) is part of a regulatory network controlling glucosinolate biosynthesis in Arabidopsis. *Plant Journal* 47 (1): 10-24

2005

Birkemeyer C, Luedemann A, Wagner C, Erban A, **Kopka J** (2005) Metabolome analysis: the potential of *in vivo*-labeling with stable isotopes for metabolite profiling. *Trends in Biotechnology* 23 (1): 28-33

Damiani I, Morreel K, Danoun S, Goeminne G, Yahiaoui N, Marque C, **Kopka J**, Messens E, Goffner D, Boerjan W, Boudet AM, Rochange S (2005) Metabolite profiling reveals a role for atypical cinnamyl alcohol dehydrogenase CAD1 in the synthesis of coniferyl alcohol in tobacco xylem. *Plant Molecular Biology* 59 (5): 753-769

Desbrosses GG, **Kopka J**, Udvardi MK (2005) *Lotus japonicus* metabolic profiling. Development of gas chromatography-mass spectrometry resources for the study of plant-microbe interactions. *Plant Physiology* 137 (4): 1302-1318

Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmueller E, Doermann P, Weckwerth W, Gibon Y, Stitt M, Willmitzer L, Fernie AR, Steinhauser D (2005) GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* 21 (8): 1635-1638

Lisso J, Steinhauser D, Altmann T, **Kopka J**, Muessig C (2005) Identification of brassinosteroid-related genes by means of transcript co-response analyses. *Nucleic Acids Research* 33 (8): 2685-2696

Lohse S, Schliemann W, Ammer C, **Kopka J**, Strack D, Fester T (2005) Organization and metabolism of plastids and mitochondria in arbuscular mycorrhizal roots of *Medicago truncatula*. *Plant Physiology* 139 (1): 329-340

Nikiforova VJ, **Kopka J**, Tolstikov V, Fiehn O, Hopkins L, Hawkesford MJ, Hesse H, Hoefgen R (2005) Systems rebalancing of metabolism in response to sulfur deprivation, as revealed by metabolome analysis of Arabidopsis plants. *Plant Physiology* 138 (1): 304-318

Schauer N, Steinhauser D, Strelkov S, Schomburg D, Allison G, Moritz T, Lundgren K, Roessner-Tunali U, Forbes MG, Willmitzer L, Fernie AR, **Kopka J** (2005) GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Letters* 579 (6): 1332-1337

Scholz M, Kaplan F, Guy CL, **Kopka J**, Selbig J (2005) Non-linear PCA: a missing data approach. *Bioinformatics* 21 (20): 3887-3895

Rautengarten C, Steinhauser D, Bussis D, Stintzi A, Schaller A, **Kopka J**, Altmann T (2005) Inferring hypotheses on functional relationships of genes: Analysis of the *Arabidopsis thaliana* subtilase gene family. *PLOS Computational Biology* 1 (4): 297-312

Urbanczyk-Wochniak E, Baxter C, Kolbe A, **Kopka J**, Sweetlove LJ, Fernie AR (2005) Profiling of diurnal patterns of metabolite and transcript abundance in potato (*Solanum tuberosum*) leaves. *Planta* 221 (6): 891-903

2004

Bino RJ, Hall RD, Fiehn O, **Kopka J**, Saito K, Draper J, Nikolau BJ, Mendes P, Roessner-Tunali U, Beale MH, Trethewey RN, Lange BM, Wurtele ES, Sumner LW (2004) Opinion: Potential of metabolomics as a functional genomics tool. *Trends in Plant Science* 9 (9): 418-425

Colebatch G, Desbrosses GG, Ott T, Krusell L, Montanari O, Kloska S, **Kopka J**, Udvardi MK (2004) Global changes in transcription orchestrate metabolic differentiation during symbiotic nitrogen fixation in *Lotus japonicus*. *Plant Journal* 39 (4): 487-512

Jenkins H, Hardy N, Beckmann M, Draper J, Smith AR, Taylor J, Fiehn O, Goodacre R, Bino RJ, Hall R, **Kopka J**, Lane GA, Lange BM, Liu JR, Mendes P, Nikolau BJ, Oliver SG, Paton NW, Rhee S, Roessner-Tunali U, Saito K, Smedsgaard J, Sumner LW, Wang T, Walsh S, Wurtele ES, Kell DB (2004) A proposed framework for the description of plant metabolomics experiments and their results. *Nature Biotechnology* 22 (12): 1601-1606

Kaplan F, **Kopka J**, Haskell DW, Zhao W, Schiller KC, Gatzke N, Sung DY, Guy CL (2004) Exploring the temperature-stress metabolome of Arabidopsis. *Plant Physiology* 136 (4): 4159-4168

Klotke J, **Kopka J**, Gatzke N, Heyer AG (2004) Impact of soluble sugar concentrations on the acquisition of freezing tolerance in accessions of *Arabidopsis thaliana* with contrasting cold adaptation. *Plant, Cell and Environment* 27(11):1395-1404

Luedemann A, Weicht D, Selbig J, **Kopka J** (2004) PaVESy: pathway visualization and editing system. *Bioinformatics* 20 (16): 2841-2844

Steinhauser D, Junker BH, Luedemann A, Selbig J, **Kopka J** (2004) Hypothesis-driven approach to predict transcriptional units from gene expression data. *Bioinformatics* 20 (12): 1928-1939

Steinhauser D, Usadel B, Luedemann L, Thimm O, **Kopka J** (2004) CSB.DB: A comprehensive systems-biology database. *Bioinformatics* 20 (18): 3647-3651

2003

Birkemeyer C, Kolasa A, **Kopka J** (2003) Comprehensive chemical derivatization for gas chromatography-mass spectrometry-based multi-targeted profiling of the major phytohormones. *Journal of Chromatography A* 993(1-2): 89-102

Hunt L, Mills LN, Pical C, Leckie CP, Aitken FL, **Kopka J**, Mueller-Roeber B, McAinsh MR, Hetherington AM, Gray JE (2003) Phospholipase C is required for the control of stomatal aperture by ABA. *Plant Journal* 34(1): 47-55

Urbanczyk-Wochniak E, Luedemann A, **Kopka J**, Selbig J, Roessner-Tunali U, Willmitzer L, Fernie AR (2003) Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Reports* 4(10): 989-993.

Wagner C, Sefkow M, **Kopka J** (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry* 62(6): 887-900

2000

Bohmert K, Balbo I, **Kopka J**, Mittendorf V, Nawrath C, Poirier Y, Tischendorf G, Trethewey RN, Willmitzer L (2000) Transgenic Arabidopsis plants can accumulate polyhydroxybutyrate to up to 4% of their fresh weight. *Planta* 211(6): 841-845

Fiehn O, **Kopka J**, Doermann P, Altmann T, Trethewey RN, Willmitzer L (2000) Metabolite profiling for plant functional genomics. *Nature Biotechnology* 18(11): 1157-1161

Fiehn O, **Kopka J**, Trethewey RN, Willmitzer L (2000) Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Analytical Chemistry* 72: 3573-3580

Roessner U, Wagner C, **Kopka J**, Trethewey RN, Willmitzer L (2000) Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant Journal* 23(1): 131-142

1998

Kopka J, Pical C, Gray JE, Mueller-Roeber B (1998) Molecular and enzymatic characterization of three phosphoinositide-specific phospholipase C isoforms from potato. *Plant Physiology* 116 (1): 239-250

Kopka J, Pical C, Hetherington AM, Mueller-Roeber B (1998) Ca^{2+}/phospholipid-binding (C-2) domain in multiple plant proteins: novel components of the calcium-sensing apparatus. *Plant Molecular Biology* 36 (5): 627-637

1997

Kopka J, Ludewig M, Mueller-Roeber B (1997) Complementary DNAs encoding eukaryotic-type cytidine-5'-diphosphate-diacylglycerol synthases of two plant species. *Plant Physiology* 113 (3): 997-1002

Kopka J, Provart NJ, Mueller-Roeber B (1997) Potato guard cells respond to drying soil by a complex change in the expression of genes related to carbon metabolism and turgor regulation. *Plant Journal* 11 (4): 871-882

1994 – 1995

Kopka J, Ohlrogge JB, Jaworski JG (1995) Analysis of in-vivo levels of acyl-thioesters with gas-chromatography mass-spectrometry of the butylamide derivative. *Analytical Biochemistry* 224 (1): 51-60

Jaworski JG, Ohlrogge JB, Tai HY, **Kopka J**, Post-Beitenmiller D (1994) Analysis of fatty-acid biosynthesis in transgenic plants with modified oil composition. *Journal of Cellular Biochemistry* 82-83 Suppl. 18A

1991 – 1993 (PhD)

Kopka J, Robers M, Schuch R, Spener F (1993) Acyl carrier proteins from developing seeds of *Cuphea lanceolata* Ait. *Planta* 191 (1): 102-111

Kopka J, Robers M, Spener F (1991) Purification and characterization of 2 acyl carrier proteins from *Cuphea lanceolata* seeds. *Biological Chemistry Hoppe-Seyler* 372 (8): 534-535

Editorials / Book chapters

Guy CL, **Kopka J**, Moritz T (2008) Plant metabolomics coming of age. *Physiologica Plantarum* 132 (2): 113-116

Hummel J, Selbig J, Walther D, **Kopka J** (2008) The Golm metabolome database: a database for GC-MS based metabolite profiling. In: Nielsen J, Jewett M (eds) Metabolomics a powerful tool in systems biology. *Topics in Current Genetics Vol. 18*, Springer-Verlag, Berlin, Heidelberg, New York, pp 75-96

Birkemeyer C, **Kopka J** (2007) Design of metabolite recovery by variations of the metabolite profiling protocol. In: Nikolau BJ, Wurtele ES (eds) Concepts in plant metabolomics. Springer-Verlag, Dordrecht, Netherlands, pp 19-38

Erban A, Schauer N, Fernie AR, **Kopka J** (2007) Non-supervised construction and application of mass spectral and retention time index libraries from time-of-flight GC-MS metabolite profiles. In: Weckwerth W (ed) Metabolomics: methods and protocols. Humana Press, Totowa, pp 19-38

Steinhauser D, **Kopka J** (2007) Methods, applications and concepts of metabolite profiling: primary metabolism. In: Fernie AR, Baginsky S (eds) Plant systems biology. *Experimentia Supplementum Vol. 97*, Verlag Birkhäuser, Basel, pp 171-194

Zuther E, Koehl K, **Kopka J** (2007) Comparative metabolome analysis of the salt response in breeding cultivars of rice. In: Jenks MA, Hasegawa PM, Jain SM (eds) Advances in molecular breeding toward drought and salt tolerant crops. Springer-Verlag, Berlin, Heidelberg, New York, pp 285-315

Kopka J (2006) Current challenges and developments in GC-MS based metabolite profiling technology. *Journal of Biotechnology* 124: 312-322

Kopka J (2006) Gas chromatography mass spectrometry. In: Saito K, Dixon RA, Willmitzer L (eds) Plant metabolomics. *Biotechnology in agriculture and forestry Vol. 57*, Nagata T, Loerz H, Widholm JM (eds), Springer-Verlag, Berlin Heidelberg New York, pp 3-20

Desbrosses GG, Steinhauser D, **Kopka J**, Udvardi MK (2005) Metabolome analysis using GC-MS. In: Marquez AJ (ed) Lotus Japonicus Handbook, Springer-Verlag, Dordrecht, pp 165-174

Kopka J, Fernie AF, Weckwerth W, Gibon Y, Stitt M (2004) Metabolite profiling in plant biology: Platforms and destinations. *Genome Biology* 5(6): 109-117

Kopka J (2004) Metabolomics: Comprehensive analysis of metabolism. In: Ganten D, Ruckpaul K (eds) Encyclopedic reference of genomics and proteomics in molecular medicine. Springer-Verlag, Berlin Heidelberg New York

Patents / Patent applications

Herold MM, Christiansen N, Kluttig M, **Kopka J**, Quedenau J. Mass spectrometry analysis method and system. PCT Int. Appl. (2006), 71pp. WO 2006082042 A2 20060810 CAN 145:198627 AN 2006:796358

Luedemann A, Erban A, Wagner C, **Kopka J**. Method for analyzing microbial metabolites by resolving isotopic mass differences within one metabolite using MALDI-TOF. PCT Int. Appl. (2005), 1050 pp. WO 2005059556 A1 20050630 CAN 143:74456 AN 2005:564800

Herold MM, Dostler M, Looser R, Walk T, Fegert A, Kluttig M, Lehmann B, Heidemann S, Hennig A, **Kopka J**. Method for extraction and analysis of contents of organic materials such as plant tissue. PCT Int. Appl. (2003), 38 pp. WO 2003041835 A1 20030522 CAN 138:398388 AN 2003:396761

Herold MM, Dostler M, Looser R, Walk T, Fegert A, Kluttig M, Lehmann B, Heidemann S, Hennig A, **Kopka J**. Single-phase mixtures for extraction of components from organic materials such as plant tissue. PCT Int. Appl. (2003), 38 pp. WO 2003041834 A1 20030522 CAN 138:398387 AN 2003:396760

Research grants

2006 -	EU FOOD-CT-2006-036220, META-PHOR "Metabolomics for plants health and outreach", in cooperation with Dr. R. Hall, Plant Research International (PRI), Wageningen, The Netherlands
2006 -	EU ERA-Net Plant Genomics 0313996B, GRASP "Research-assisted breeding for the sustainable production of quality grapes and wines. Subprojekt: GC-MS based metabolome studies" in cooperation with Dr. E. Zyprian, Prof. R. Töpfer, Institut für Rebenzüchtung (IRZ), Geilweilerhof, Germany
2006 -	BMBF QuantPro program, InnOx "Innovative diagnostic tools to optimise potato breeding: systemic analysis of cellular processes and their relation to plant internal oxygen concentrations", in cooperation with Dr. P. Geigenberger, Dr. J. van Dongen, Department Prof. M. Stitt, MPIMP, Potsdam-Golm
2006 -	DFG KO2329/3-1, "Influence of inorganic carbon and mutations in carbon-regulated pathways on metabolite pools and turnover in cyanobacteria" in cooperation with PD Dr. M. Hagemann, University of Rostock, FB Biosciences, Department of Plant Physiology, Germany
2005 -	EU INCO-CT-2005-517617, LOTASSA "Bridging genomics and agrosystem management: Resources for adaptation and sustainable production of forage Lotus species in environmentally-constrained south-american soils" in cooperation with Dr. M.K. Udvardi, MPIMP, Potsdam-Golm
2002 - 2007	BMBF 0312854, "Improving the quality of rice with respect to abiotic stress resistance and nutritional parameters", in cooperation with the Institute of Biotechnology, Hanoi, Vietnam
2001 - 2004	PhD grant, "Metabolite identification in GC-MS profiles of *Arabidopsis thaliana* roots", Metanomics GmbH & CoKG aA, Berlin
1989 - 1991	PhD stipend, Graduiertenförderung of the state ministry of science and research, Northrhine-Westfalia, Germany

University teaching record

WS 2006/2007	Joachim Kopka, „Vergleichende Metabolom Analyse niederer Eukaryonten, Prokaryonten und höherer Pflanzen (Mikrobiologie)", Vorlesung/Seminar DB/DBC Hauptstudium (2 SWS), ab 16.10.2006 Praktikum DB/DBC Hauptstudium 2 Wochen ganztägig (6 SWS), nach Vereinbarung, (Mikrobiologie, max. 10 Teilnehmer), Teilnahme an S wird vorausgesetzt,

	Vorlesungs- und Personalverzeichnis WS 2006/2007 der Universität Potsdam
WS 2005/2006	Joachim Kopka, „Vergleichende Physiologie niederer Eukaryonten, Prokaryonten und höherer Pflanzen (Mikrobiologie)“, Vorlesung/Seminar DB/DBC Hauptstudium (2 SWS), ab 17.10.2005 Praktikum DB/DBC Hauptstudium 2 Wochen ganztägig (6 SWS), nach Vereinbarung, (Mikrobiologie, max. 10 Teilnehmer), Teilnahme an S wird vorausgesetzt, Vorlesungs- und Personalverzeichnis WS 2005/2006 der Universität Potsdam
WS 2004/2005	Joachim Kopka, „Vergleichende Physiologie niederer Eukaryonten, Prokaryonten und höherer Pflanzen (Mikrobiologie)“, Vorlesung/Seminar DB/DBC Hauptstudium (2 SWS), ab 11.10.2004 Praktikum DB/DBC Hauptstudium 2 Wochen ganztägig (2,5 SWS), nach Vereinbarung, (Mikrobiologie, max. 10 Teilnehmer), Teilnahme an S wird vorausgesetzt, Vorlesungs- und Personalverzeichnis WS 2004/2005 der Universität Potsdam

Invited lectures

2009	January 27th-28th, “Standardized GC-MS based metabolome data for the Golm Metabolome Database (GMD)”, International Metabolomics Workshop, European Bioinformatics Institute (EBI), Cambridge, UK
2008	October 7th-9th, “Qualitätsverbesserung von Reis bezüglich der Resistenz gegen abiotischen Stress und ernährungsphysiologischer Eigenschaften”, Biotechnica 2008, 2. BMBF-Projektforum Biotechnologie, Hannover, Germany
2008	September 7th-12th, “Advances in metabolite phenotyping: Exploration of the metabolic complement of the salt stress acclimation in plants”, Gordon Research Conference (GRC): Salt and water stress in plants, Big Sky, MT, USA
2008	September 2nd-5th, “Mass spectrometry in metabolomics: Status and future development of GC-TOF-MS metabolite profiling”, Nutrigenomics (NuGO) Week, Potsdam, Germany
2008	June 9th-10th, “Metabolic adaption of plants under stress”, Trends in Metabolomics (DECHEMA e.V.), Frankfurt am Main, Germany
2008	May 24th-26th, “GC-MS analysis of rice: An introduction: Metabolite profiling applied to assess the metabolic component of rice salt acclimatization”, Metabolomics and Rice Quality, Vientiane, Laos
2008	May 16th-18th, “Tools for metabolic phenotyping and their contribution to systems biology”, Nobel Conference on Systems Biology and Child Health, Stockholm, Sweden
2007	December 11th-15th, “Data processing for automated GC-MS based metabolite profiling”, Joint 80th Annual Meeting of the Japanese Biochemical Society and the 30th Annual Meeting of the Molecular Biology Society of Japan (BMB 2007), Yokohama Pacifico, Japan
2007	September 25th-26th, “GC-TOF-MS Metabolite profiling: Recent developments and applications”, BIC-GH Bioinformatic Symposium, Halle, Germany
2007	September 5th-7th, “GC-EI-TOF-MS Analysis of *in vivo*-carbon-partitioning into soluble metabolite pools”, Workshop Molecular Interactions, Berlin, Germany
2007	June 12th-14th, “GC-EI-TOF-MS Analysis of *in vivo*-carbon-partitioning into soluble metabolite pools of higher plants by monitoring isotope dilution after $^{13}CO_2$ labeling”, 3rd Annual International Conference of the Metabolomics Society, Manchester, UK
2007	June 11th-12th, “GCxGC-TOF-MS: Discussion of potential and limitations for metabolite profiling”, Workshop of the 3rd Annual International Conference of the Metabolomics Society, Manchester, UK
2006	January 24th -27th, “Challenges and developments in GC-EI-MS based metabolite profiling: Enhanced metabolite identification and metabolome characterization”, 4th BIORHIZ International Workshop: Rhizosphere processes and induced defense, Jena Germany
2005	September 25th-29th, “Challenges and developments in GC-EI-MS based metabolite profiling: Enhanced metabolite identification and metabolome characterization”, ComBio2005, Adelaide, Australia

2005	August 1st-2nd, "Challenges and developments in GC-EI-MS-based metabolite profiling: Enhanced metabolite identification and metabolome characterization", Metabolomics Standards Workshop, Bethesda, MA, USA
2005	June 20th-23rd, "Mass spectral and retention index libraries for GC-EI-MS based metabolome studies: Enhanced metabolite identification and characterization in complex biological profiles", 1st Annual International Conference of the Metabolomics Society, Tsuruoka City, Japan
2005	May 10th-12th, "Current Challenges and developments in metabolite profiling", BioPerspectives 2005, Wiesbaden, Germany
2005	April 27th-30th, "Current challenges and perspectives of plant metabolomics", Genomics and Beyond: Frontiers in Plant Biology, Columbia, MO, USA
2005	February 15th, "Tools to gain insight from metabolite profiles", Symposium in Metabolomics, Umea, Sweden
2004	November 5th-7th, "Metabolic profiling: A tool to gain insight into the metabolism", 7th Nordic Photosynthesis Congress, Turku, Finland
2004	September 21st-25th, "Metabolomics: A tool to gain insight into gene and metabolite function", 3rd Plant Genomics European Meeting (Plant GEMs), Lyon, France
2004	June 3rd-6th, "Mass spectral libraries for metabolite identification and characterization in complex GC-EI-MS profiles", 3rd International Congress on Plant Metabolomics, Ames, IA, USA
2004	May 7th-16th, "Metabolic profiling: a tool to gain insight into the metabolome", Workshop Systems Biology – From Genome to Phenome, La Trobe University, Melbourne, Australia
2003	September 28th-October 1st, "Metabolite profiling a potential tool to establish substantial equivalence of GMO material", Gene Flow Conference, Mexico City, Mexico
2003	June 23rd-28th, "Metabolic profiling: A tool to gain insight into the plant metabolome", 7th International Congress of Plant Molecular Biology (ISPMB), Barcelona, Spain
2003	May 19th-24th, "Co-response analysis of gene expression and metabolite profiles for the discovery of functional units", 27th International Exhibition-Congress on Chemical Engineering, Environmental Protection and Biotechnology at the ACHEMA 2003, Frankfurt am Main, Germany
2003	April 25th-28th, "Application of GC-TOF-MS technology to metabolic profiling", 2nd International Conference on Plant Metabolomics, Potsdam, Germany
2002	May 13th-14th, „Bedeutung multiparalleler Metabolitenanalysen für die biologische Systemanalytik", 4. BMBF Biotechnologie-Tage, Braunschweig, Germany

Conference/ Workshop organization

2008	December 10th-12th, organizer, Metabolome Data Processing Workshop, Potsdam-Golm, Germany
2008	July 15th-18th, member of the advisory board, 5th International Conference on Plant Metabolomics, Pacifico Yokohama, Japan
2008	June 9th-10th, member of the organizing committee, Trends in Metabolomics (DECHEMA e.V.), Frankfurt am Main, Germany
2007	April 23rd-27th, co-organizer, Workshop "Profiling Technologies & Bioinformatics", Potsdam-Golm, Germany
2006	September 20th-24th, co-organizer, PlantMetaNet ETNA metabolomics research school on "Signals, Sensing and Plant Primary Metabolism", Potsdam-Golm, Germany
2006	April 7th-10th, member of the advisory board, 4th International Conference on Plant Metabolomics, Reading, Berkshire, UK
2003	April 25th-28th, co-organizer, 2nd International Conference on Plant Metabolomics, Potsdam, Germany

RELEVANT PUBLICATIONS

Appendix A: Metabolite Profiling: Concepts, Basic Method Descriptions, Analytical Technology Enhancement

[1] Fiehn O, **Kopka J**, Doermann P, Altmann T, Trethewey RN, Willmitzer L (2000) Metabolite profiling for plant functional genomics. *Nature Biotechnology* 18(11): 1157-1161
(http://dx.doi.org/10.1038/81137)
(http://www.nature.com/nbt/journal/v18/n11/abs/nbt1100_1157.html)

[2] Roessner U, Wagner C, **Kopka J**, Trethewey RN, Willmitzer L (2000) Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant Journal* 23(1): 131-142
(http://dx.doi.org/10.1046/j.1365-313x.2000.00774.x)
(http://www3.interscience.wiley.com/journal/119188238/abstract?CRETRY=1&SRETRY=0)

[3] Lisec J, Schauer N, **Kopka J**, Willmitzer L, Fernie AR (2006) Gas chromatography mass spectrometry-based metabolite profiling in plants. *Nature Protocols* 1: 387 – 396
(http://dx.doi.org/10.1038/nprot.2006.59)
(http://www.nature.com/nprot/journal/v1/n1/abs/nprot.2006.59.html)

[4] Erban A, Schauer N, Fernie AR, **Kopka J** (2007) Non-supervised construction and application of mass spectral and retention time index libraries from time-of-flight GC-MS metabolite profiles. In: Weckwerth W (ed) Metabolomics: methods and protocols. Humana Press, Totowa, pp 19-38
(http://dx.doi.org/10.1007/978-1-59745-244-1_2)
(http://www.springerlink.com/content/v665123215807178/)

[*] *First experimental demonstration of feasibility, proof of concept, and establishment of the basic workflow of GC-MS based metabolite profiling studies leading to the discovery of the wealth of identified and also yet non-identified metabolic components accessible by this technology.* My achievements in the initial phases of the emergence of the metabolomics field were both, conceptual and experimental. I performed the initial method establishment for relative quantitative chemical analysis by GC-MS, initiated compound identification using mass spectral and chromatographic retention index libraries as part of the chromatography data processing and established first data mining and statistical approaches of metabolite profiles. The basic workflow was subsequently refined and applied to specific aspects of plant physiology by the main authors of the above publications. In the years 1998-2000 I contributed the transfer of the metabolite profiling concept to commercial applications as a founding member and head of the internal feasibility study of the Metanomics GmbH & Co. KGaA. This technology transfer was documented by PCT Int. Appl. WO 2006082042 A2 20060810, WO 2003041835 A1 20030522, and WO 2003041834 A1 20030522.

[5] Birkemeyer C, Kolasa A, **Kopka J** (2003) Comprehensive chemical derivatization for gas chromatography-mass spectrometry-based multi-targeted profiling of the major phytohormones. *Journal of Chromatography A* 993(1-2): 89-102
(http://dx.doi.org/10.1016/S0021-9673(03)00356-X)
(http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6TG8-48BKHH0-B&_user=10&_coverDate=04%2F18%2F2003&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=93385d00191686e7fd233ea549f2a19a)

[*] Footnotes indicate my personal contributions to the above publications. Here, publications are sorted by research topics rather than alphabetically (cf. REFERENCES) or chronologically (cf. CURRICULUM VITAE).

Project leadership of the exploration of the tool box of chemical derivatization reactions for applications in metabolite and future trace compound profiling. This study was central part of the PhD thesis of Dr. Claudia Birkemeyer.

[6] Birkemeyer C, Luedemann A, Wagner C, Erban A, **Kopka J** (2005) Metabolome analysis: the potential of *in vivo*-labeling with stable isotopes for metabolite profiling. *Trends in Biotechnology* 23 (1): 28-33 (http://dx.doi.org/10.1016/j.tibtech.2004.12.001) (http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6TCW-4F1J8P7-1&_user=10&_coverDate=01%2F01%2F2005&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=714b5ff6f52082d0d68e26821435b76b)

Project leadership of the development of in-vivo stable isotope labeling for the purpose of multiplexed internal quantitative standardization. Enhancing the precision of relative quantification is seen as a prerequisite for extending metabolite profiling studies towards enrichment and sub-fractionation of trace compounds. Prior to publication this concept and respective feasibility studies using the *Saccharomyces cerevisiae* model were filed as PCT Int. Appl. WO 2005059556 A1 20050630.

[7] Huege J, Sulpice R, Gibon Y, Lisec J, Koehl K, **Kopka J** (2007) GC-EI-TOF-MS analysis of *in vivo*-carbon-partitioning into soluble metabolite pools of higher plants by monitoring isotope dilution after ($^{13}CO_2$)-labelling. *Phytochemistry* 68 (16-18): 2258-2272 (http://dx.doi.org/10.1016/j.phytochem.2007.03.026) (http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6TH7-4NMCV7Y-3&_user=10&_coverDate=09%2F30%2F2007&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=176c293e2f839f93c7cc8fba00412a24)

Project leadership of method development towards dynamic flux studies of higher plants. The full ^{13}C-labelling of the higher plants, namely, *Arabidopsis thaliana* and *Oryza sativa*, was established and a first assessment of assimilated CO_2-partitioning into metabolite pools of roots and leaves performed under controlled environmental conditions. This work was central to the diploma thesis of Jan Huege.

Metabolite profiling for plant functional genomics

Oliver Fiehn[1]*, Joachim Kopka[2], Peter Dörmann[1], Thomas Altmann[1], Richard N. Trethewey[2], and Lothar Willmitzer[1]

[1]Max Planck Institute of Molecular Plant Physiology, 14424 Potsdam, Germany. [2]Metanomics GmbH & Co KGaA, Tegeler Weg 33, 10589 Berlin, Germany.
*Corresponding author (fiehn@mpimp-golm.mpg.de).

Received 17 April 2000; accepted 21 August 2000

Multiparallel analyses of mRNA and proteins are central to today's functional genomics initiatives. We describe here the use of metabolite profiling as a new tool for a comparative display of gene function. It has the potential not only to provide deeper insight into complex regulatory processes but also to determine phenotype directly. Using gas chromatography/mass spectrometry (GC/MS), we automatically quantified 326 distinct compounds from *Arabidopsis thaliana* leaf extracts. It was possible to assign a chemical structure to approximately half of these compounds. Comparison of four *Arabidopsis* genotypes (two homozygous ecotypes and a mutant of each ecotype) showed that each genotype possesses a distinct metabolic profile. Data mining tools such as principal component analysis enabled the assignment of "metabolic phenotypes" using these large data sets. The metabolic phenotypes of the two ecotypes were more divergent than were the metabolic phenotypes of the single-loci mutant and their parental ecotypes. These results demonstrate the use of metabolite profiling as a tool to significantly extend and enhance the power of existing functional genomics approaches.

Keywords: functional genomics, *Arabidopsis thaliana*, metabolite profiling, cluster analysis, metabolomics, bioinformatics

In the post-genomic era, elucidation of gene function is a main focus. Plant functional genomics[1] couples the generation of transgenic and mutant plants to the multiparallel analysis of gene products such as mRNA[2] and proteins[3]. However, these methods do not provide direct information about how a change in mRNA or protein is coupled to a change in biological function. As a result of a multiplicity of regulatory interactions at all levels in plant cells, a change at one level in the complex network does not necessarily lead to a particular change in function or phenotype. Instead, single point mutations might often lead to complex responses at the level of the whole organism. In applying the profiling concept, it is crucial to perform unbiased (metabolite) analyses in order to define precisely the biochemical function of plant metabolism[4]. Such analyses complement existing functional genomics methodologies while offering a direct link between a gene sequence and the function of the metabolic network in plants. Furthermore, metabolite profiling can elucidate links and relationships that occur primarily through regulation at the metabolic level. Finally, a broad metabolic analysis may address public concerns about the safety and value of plant genetically modified organisms.

To become established as a robust tool, metabolite profiling must be fast, reliable, sensitive, and suitable for automation, as well as covering a significant number of metabolites. A range of analytical technologies enhances the sensitivity and universality of mass spectrometry by chromatographic separations. To date, however, metabolic screening approaches using mass spectrometry are rarely used in plant research[5,6]. For the most part, the use of multitarget profiling has been limited to rapid clinical detection of human diseases[7]. We judged gas chromatography coupled to electron-impact quadrupole mass spectrometry (GC/MS) to be the most mature technology capable of fulfilling the required criteria. The methodology described here allows the detection and quantification of more than 300 compounds from a single plant leaf extract.

Results and discussion

Plant leaf extracts yield 326 quantifiable compounds. Metabolite extraction from *Arabidopsis* leaf tissue was done using methanol and heat, thereby rapidly inhibiting enzymatic activity. We added internal standards in order to correct for minor variations occurring during sample preparation and analysis. A single fractionation step into a lipophilic and a polar phase was followed by solvent evaporation and derivatization for increasing metabolite stability and volatility as reported[8]. Briefly, the lipid phase was transmethylated and trimethylsilylated for the analysis of total fatty acids, fatty alcohols, sterols, and aliphatics, whereas the polar phase was methoximated and trimethylsilylated for the analysis of hydroxy- and amino acids, sugars, sugar alcohols, organic monophosphates, (poly)amines, and aromatic acids. Metabolite sizes were in the range of ethylene glycol (62 AMU) to trisaccharides (504 AMU). Optimal reaction conditions were established as a compromise between reaction completeness and the maintenance of labile compound integrity (data not shown). We chose analytical parameters as a compromise between separation efficiency, column capacity, and column long-term stability. This GC/MS approach is extremely powerful for plant metabolite profiling (Fig. 1). Hundreds of different compounds were detected in parallel, some of which had severely overlapping peaks that are deconvoluted by selective ion traces (Fig. 1B). Compound identification was performed by comparison of mass spectra and retention times with those obtained with commercially available reference compounds. A major advantage of mass spec-

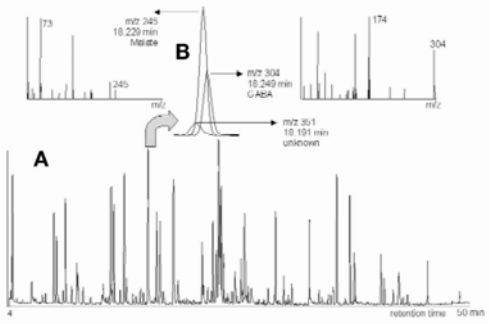


Figure 1. Metabolite profiling by GC/MS. Base peak intensity GC/MS chromatogram of the polar fraction of a leaf extract from the *Arabidopsis dgd1* mutant (A). Target metabolites are identified by exact retention times and their corresponding mass spectra (B) as shown for the co-eluting peaks of malate, γ-aminobutyric acid (GABA), and an unidentified compound. m/z, Ratio of mass to charge.

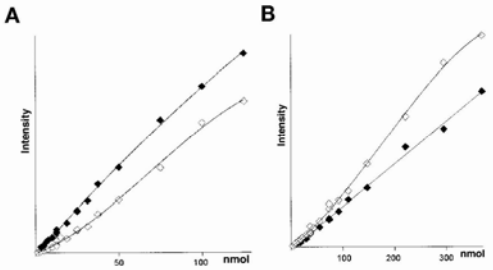


Figure 2. Metabolite calibrations. Calibration curves for determination of dynamic ranges and absolute concentrations using stable isotope-labeled metabolites. Open symbols, external calibration; filled symbols, internal calibration. (A) $^{13}C_6$-Glucose. (B) d_4-ethanolamine.

trometry is that unknown peaks can be determined as reliably as known target analytes without prior knowledge of their exact chemical structure. In *Arabidopsis* extracts, after rigorous comparison of mass spectra with commercial libraries[9], about half of the detected peaks currently remain unidentified. High-throughput peak finding was done by matching mass spectra within a 0.25 min wide time window around the predicted retention times for each target compound. Fluctuations in the relative retention times of sugars and hydroxy acids were found to lie within 0.02 min of the predicted time and, thus, occasional false positive identifications were corrected by setting a postacquisition threshold for the deviation of the retention time of 0.04 min (0.07 min threshold for amino compounds). False positive identifications were automatically qualified as not determined and excluded from further calculations or manually corrected, where necessary. In total, 326 compounds were found in the *Arabidopsis thaliana* leaf extracts (101 polar and 63 lipophilic identified compounds, plus 113 polar and 49 lipophilic compounds of unknown chemical structure). A complete list of the mass spectra of our current target compounds and sample preparation protocols can be downloaded from our website[10].

With respect to quantification, we followed two approaches. Relative amounts of the various compounds were obtained by normalizing the intensity of individual ion traces (that are indicative for the respective compound even in the presence of co-eluting compounds) to the response of internal reference compounds, and further, to 1 mg of plant leaf fresh weight. For quantification a linear relationship between metabolite amount and the analytical signal is crucial. Internal calibration curves confirmed that this assumption holds true over two to three orders of magnitude when 11 stable isotopic labeled compounds were added to 32 different *Arabidopsis thaliana* leaf extracts (Fig. 2). Because of matrix effects, up to twofold differences were found between external and internal calibration, but no differences were found between mutant and wild-type C24 plants. Calibration linearity was also confirmed for 50 metabolites of unidentified chemical structure both by diluting derivatized plant samples and by derivatizing different volumes of a single plant extract (data not shown).

The stable isotope internal calibration curves were also used to determine the absolute amounts of certain metabolites. Table 1 summarizes the absolute mean values for these compounds as determined for leaf extracts of 18 individual *Arabidopsis thaliana* C24 wild-type plants. Graphs in Figure 2 and data contained in Table 1 confirm that the profiling method established here allows the determination of both relative and absolute quantities. However, it is important to stress that for the vast majority of applications of the profiling technology, the absolute value is unimportant, rather the relative value is sufficient.

Another key factor for any analytical technique is reproducibility. The reproducibility of the whole process was tested in order to determine the potential contribution of variability in the analytical method to the observed variation between different biological samples. In order to estimate the influence of the sample preparation and the analytical device on variability, two samples of *Arabidopsis* C24 wild type were combined directly following extraction and divided into seven aliquots. Each aliquot was taken separately through the sample preparation procedure and after GC/MS analysis of the polar phase, relative standard deviations were determined for 149 polar metabolites. The mean of these deviations was 8% ± 6%, and 110 of these compounds showed even lower deviations (5% ± 2%). This is at least as accurate as comparable functional genomic methods at the protein level[11] and clearly more accurate than differential analysis of expression using cDNA microarrays[12]. Therefore, we conclude that the variability introduced into the analytics by the sample preparation and the actual measurement is small and can be tolerated.

In order to get an insight into the biological variability, 18 plants of *Arabidopsis thaliana* genotype C24 (wild type) were grown in the phytotron side by side under identical conditions and harvested at the same time. Absolute values were determined for 11 metabolites based on isotope-labeled internal calibration curves. As evident from Table 1, the variability due to the biological variability is in clear excess of the variability due to the overall analytical precision. This finding indicates the metabolic flexibility of plants. For the

Table 1. Biological variation and analytical precision in *Arabidopsis thaliana* C24 WT plants

Chemical	Average contents (nmol/mg FW)[a], *n* = 18	Biological variation (% s.d.), *n* = 18	Analytical reproducibility (% s.d.), *n* = 7
Ethylene glycol	1.2 ± 0.3	26	6
Alanine	144 ± 56	38	12
Valine	38 ± 6	17	5
Ethanolamine	63 ± 14	23	6
Glycerol	24 ± 12	49	2
Leucine	25 ± 7	29	8
Benzoic acid	0.8 ± 0.3	40	10
Aspartic acid	79 ± 34	43	5
Glutamic acid	199 ± 75	38	3
Glucose	119 ± 67	56	5
Sucrose	598 ± 180	30	2

[a]FW, fresh weight.

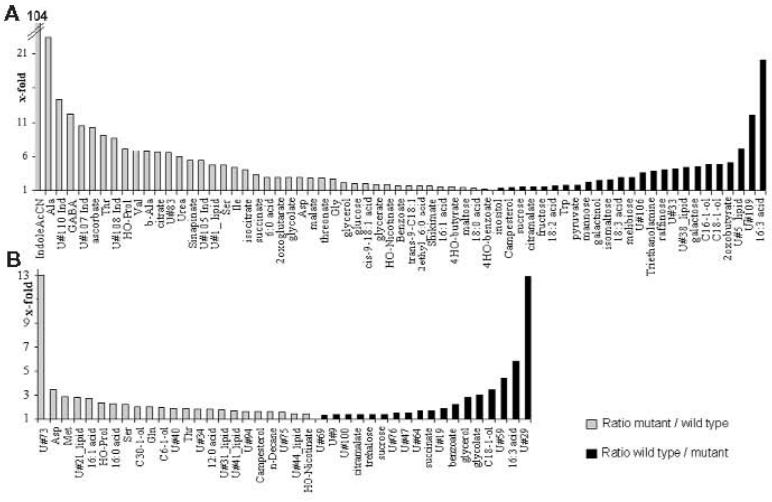


Figure 3. Significant metabolite differences in plant genotypes. Alterations in mean metabolite levels (*t*-test, $p < 0.01$) of (A) the *dgd1* mutant and (B) the *sdd1-1* mutant compared to their respective parental wild-type backgrounds (Col-2 and C24). For the *dgd1* mutant, 153 significant alterations in metabolites were found. For visual clarity, only 67 of the metabolites are presented that were selected either by their physiological importance or by their metabolite alteration exceeding a factor of 3. For *sdd1-1*, all 41 of the significant alterations are shown.

remaining 140 compounds only relative quantifications were performed. Again the biological variability was found to be on average ~40% s.d. These findings therefore show that the biological variability seen between genetically identical plants grown under identical conditions is the largest source for the variability observed.

For application in the framework of genomic approaches, one single individual can process 60 samples per working day. Using our protocols, three GC/MS machines are needed for the processing of the 60 samples. It is obvious that this figure can be easily amplified by increasing the number of machines and persons involved.

Mutants and parental ecotypes display large metabolic differences. The power of metabolite profiling was tested for its ability to distinguish between ecotypes using various *Arabidopsis thaliana* genotypes, which are supposedly genetically characterized by the presence of several hundred allelic differences, and two mutants with these ecotype backgrounds. One of the mutants should display a severe visible phenotype, whereas the other mutant should grow and develop essentially indistinguishable from the parental wild-type background. The two ecotypes chosen by us were Col-2 and C24. On the genetic background of Col-2, the *dgd1* mutant was chosen, which is characterized by a 90% reduction in the galactolipid digalactosyldiacylglycerol (DGD)[13]. As a consequence of the reduced levels of DGD, the mutant is impaired in photosynthesis and is hypersensitive to light stress[14], and thus served as an example of a rather severe phenotype. The gene affected was recently cloned and shown to encode a galactosyl transferase (DGD synthase). Because the mutant was backcrossed four times with the parental ecotype, Col-2, most of the original mutant DNA was replaced by Col-2 DNA. By transformation of this line with wild-type genomic DNA fragments carrying the *DGD1* gene or with the DGD1 cDNA, we could demonstrate[15] that not only the DGD lipid phenotype but also the growth defect were complemented. Therefore, all effects other than deficiency in DGD biosynthesis are believed to be secondary effects. The second mutant used in this study, *sdd1-1*, carries a point mutation in a regulatory gene involved in the control of stomatal development[16]. Like *dgd1*, *sdd1-1* was also backcrossed four times with its parental ecotype, C24. The lack of *SDD1* gene function causes a two- to fourfold increase in stomatal density; however, the mutant displays no other visible phenotype, and therefore was chosen to represent a mild mutant phenotype. Thus, *sdd1-1* was selected as a morphological mutant for analysis to gain information about the potential metabolic changes caused by the increased stomatal density that result in enhanced gas exchange properties (increased CO_2 uptake and H_2O release) of the leaves.

Mutant plants were grown in parallel with their corresponding wild-type plants until the flowering stage (defined by the presence of an inflorescence stem about 7 cm in height) in a controlled environment under standard conditions. All plants were randomly distributed within the growth chamber to eliminate a potential contribution of position effects. For analysis of each genotype samples from fully expanded rosette leaves were taken from 28–45 individual plants. Individual processing of these samples resulted in 28–45 individual profiles per genotype. After GC/MS analysis, data normalization, and data validation, Student's *t*-tests were carried out for statistical analysis[17]. To achieve high result reliability, we used *t*-test probability limits of $p < 0.01$ in our evaluations.

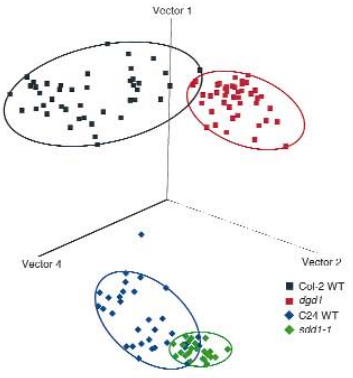


Figure 4. Metabolic phenotype clustering. Clusters found after principal component analysis (PCA) of log-scaled polar metabolite data of 151 samples originating from four plant genotypes. Single-loci mutants show metabolic phenotypes distinct from wild-type plants (WT). Basic vectors in PCA span an *n*-dimensional space to give best sample separation. Each point represents a linear combination of all the metabolites from an individual sample. Vectors 1, 2, and 4 were chosen for best visualization of genotype separation and include 62% of the total information content derived from metabolite variances.

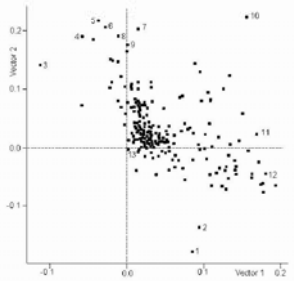


Figure 5. Metabolite impacts on clustering results. The contribution of individual polar metabolites to the PCA vector calculation is computed by linear combination. The closer to zero, the less influence of a metabolite on linear combination is found. Vector 1 predominantly separates plants from C24 and Col-2 genetic backgrounds, whereas vector 2 separates Col-2 ecotype plants from the corresponding *dgd1* mutants. Vector 4 contributed most to the separation of C24 wild-type plants from the *sdd1-1* mutants, although for reasons of clarity this vector is not shown in this figure. Examples of metabolite identity are numbered: 1 = isomaltose; 2 = U#106; 3 = proline; 4 = serine; 5 = threonine; 6 = pyroglutamic acid; 7 = glutamate; 8 = β-alanine; 9 = phenylalanine; 10 = U#107 (indole derivative); 11 = ascorbate; 12 = γ-hydroxybutyric acid lactone;13 = U#72.

The loss of activity of a single enzyme in the *dgd1* mutant resulted in a dramatic alteration in the metabolite composition: in comparison to the corresponding Col-2 wild-type plants, the levels of 153 out of 326 quantified metabolites were significantly different (Fig. 3A) in the *dgd1* mutant plants. The metabolic differences between Col-2 wild type and the *dgd1* mutant are quite complex and at present can only be partially explained. For example, some amino acids and citrate cycle intermediates are increased in the *dgd1* mutant, possibly indicating an increase in citrate cycle activity. Furthermore, indole-3-acetonitrile and several unidentified indole derivatives were increased in the mutant. Indole-3-acetonitrile is the precursor of the plant hormone indole-3-acetic acid (IAA), which itself did not reach detectable amounts by our profiling approach. The differences in IAA metabolism may reflect a hormone-controlled mechanism induced by the growth retardation of the *dgd1* mutant. Concomitant with the reduction in DGD lipid, the amount of the fatty acid 16:3 is decreased in the mutant, which can be explained by a change in the relative amounts of different forms of the substrate of the DGD synthase reaction, monogalactosyl diacylglycerol. The apparent reduction of galactose content in the *dgd1* mutant may reflect the downregulation of overall galactose biosynthesis as a response to the block in galactolipid biosynthesis. Furthermore, the concomitant reductions of inositol, galactinol, raffinose, and melibiose point toward a reduced flux through the biosynthesis of the carbohydrates of the galactinol family. It is obvious from its metabolic profile that a wide range of enzymes and pathways have been affected by the *dgd1* mutation. This analysis demonstrates the power of the metabolite profiling method to identify and quantify previously overlooked alterations, allowing a more comprehensive interpretation of the consequences of genetic modifications.

The second mutant that we tested, *sdd1-1*, is deficient in a subtilisin-like serine protease likely to be involved in the processing of a proteinaceous component of a signal transduction pathway controlling stomatal development. Apart from exhibiting increased stomatal density and stomatal cluster development, the *sdd1-1* mutant does not display any other obvious visible morphological alterations. However, it does have a slight retardation in seedling establishment after sowing, which results in a three- to four-day delay in flowering under the conditions used. In contrast to the drastic alterations observed in the *dgd1* mutant, there were fewer variations in metabolite levels in the *sdd1-1* mutant when compared to the corresponding wild type. Significant differences were found in 41 metabolites (Fig. 3B), but only a few compounds were altered more than twofold. None of these changes are obviously linked to the elevated stomatal density in *sdd1-1*. Metabolite levels were expected to be increased for osmotically active components (as compensatory reactions to elevated transpiration) or carbohydrates (because of a raised net CO_2 uptake mediated by the enhanced stomatal density). Metabolite profiling, however, revealed neither a net increase in osmolytes nor in primary products of photosynthesis. The most dramatic difference between the *sdd1-1* mutant and the wild type occurred for two hydrophilic substances of unknown identity. The 13-fold reduction in substance U#29 and the concomitant 13-fold increase in U#73 may be indicative of a close metabolic relationship. The effects of the *sdd1-1* mutation on lipophilic metabolites are as difficult to understand as the alteration in polar metabolism. There is a significant change in leaf fatty acid composition: One of the most abundant fatty acids in *Arabidopsis*, 16:3, is decreased by more than fivefold in mutant plants, whereas 16:1 and 16:0 are increased and most of the other fatty acids remained unaffected. It has been shown recently by means of gene silencing that decreases in 16:3 levels lead to an improved thermo tolerance in transgenic plants[18] by modification of membrane function. Therefore, it is an interesting finding that the single-loci mutations tested here, *sdd1-1* and *dgd1*, also lead to a decrease of 16:3.

In both mutants alterations in many metabolites were observed. As stated above, half of the scored metabolites are of unknown structure. Because metabolite profiling reveals in cases such as *sdd1-1* that the most dramatic changes occur in unknown metabolites, further analyses, including structure elucidation, can be focused on a small number of compounds. In addition, new plant metabolites from unknown pathways can be detected by a non-target profiling approach. Triethanolamine is not commonly known as a plant endogenous metabolite in standard biochemical pathways[19,20] but is of widespread use as an organic solvent. The fact that significantly decreased levels of triethanolamine were observed in *dgd1* plants compared to Col-2 wild-type plants strongly argues against it being a contaminant, and rather suggests that it is produced by the plant biosynthetic machinery.

Principal component analysis reveals four clusters. Data interpretation of mean metabolite levels is difficult not only because biochemical pathways are linked and highly regulated but also because information gets lost in the process of averaging. Each individual plant represents a unique biological system; thus, it is to be expected that metabolite correlations to gene functions will be more clearly distinguished by multivariate data mining techniques[21]. Data mining tools reduce data complexity by focusing on the information content of a given data set. Two methods were applied: hierarchical component analysis (HCA) and principal component analysis (PCA)[22]. Both methods use all metabolite data from a plant sample to compute an individual metabolic profile and simultaneously compare this profile with all other plant metabolic profiles. As a first example, calculation of pattern recognition was based on the metabolic profiles of the polar compounds. In HCA, this pattern recognition is performed by calculation of Euclidean distances resulting in groups of samples (clusters) that show multivariate similarity. By examining the corresponding HCA dendrogram we found two main clusters for each of the *Arabidopsis* ecotypes. Each of these clusters was further divided into two subclusters corresponding to wild-type and mutant plants (data not shown). This genotype clustering was confirmed by PCA pattern recognition, which in some ways is an even more useful approach for the identification of gene function from metabolic profiles. By an *n*-dimensional

vector approach, PCA finds those basic vectors (eigenvectors) that give best overall sample separation. On the basis of total variances, vectors are determined by linear combination of all metabolite data. The resulting vectors are ordered by decreasing amount of total variance resulting in a minimum of loss of information content when data are visualized. Each sample can then be represented in a two- or three-dimensional space spanned by these vectors. When all samples of a genotype accumulate in the same cluster, this cluster can be regarded as a specific "metabolic phenotype." After application of PCA algorithms to the Col-2 / *dgd1* / C24 / *sdd1-1* experimental data set of polar compounds, four different clusters were found that are identical with the four plant genotypes (Fig. 4). For visualization, vector 4 was chosen instead of vector 3, which had nearly the same information content but was less powerful in separating C24 WT from *sdd1-1* samples. Plants with the Col-2 genetic background were quite dissimilar from C24 plants, whereas the difference between the two wild types and their corresponding mutant metabolic phenotypes were smaller or even partially overlapping (C24 WT / *sdd1-1*). This finding corresponds well to the results obtained from Student's *t*-tests of individual metabolites, where metabolite differences were both more abundant and more extreme for the Col2 WT / *dgd1* samples when compared with the C24 WT / *sdd1-1* samples.

Furthermore, PCA data can be used to analyze which metabolites exert the largest influence on the basic vector calculation (Fig. 5). For example, for computing the most powerful PCA vectors 1 and 2, many metabolites had values near zero, indicating that only minor variances were observed. However, some metabolites such as isomaltose, unknown #106, serine, threonine, β-alanine, and the unknown indole derivative #107 had a comparatively strong impact on the calculation of PCA vector 2, which separated predominantly Col-2 WT from *dgd1* plants. These compounds also demonstrated $p < 0.01$ in the *t*-test comparison. Additionally, PCA vector 2 was strongly influenced by metabolites that were not significantly different in *t*-tests, either because these metabolites did not match the *t*-test threshold (pyroglutamic acid, $p = 0.015$; phenylalanine, $p = 0.048$; glutamate, $p = 0.022$) or because they were not detectable in one of the two genotypes being compared, which causes *t*-tests to fail (proline, ascorbate). Analysis of PCA vector loading supplies information for the interpretation of metabolic profiles that extends the results obtainable by classical *t*-tests. For ease of visualization, vector 4 was left out in this presentation.

The ability to assign plant samples to groups using PCA of metabolic profiles offers an exciting perspective for plant functional genomics. On one hand, such groups are likely to be defined predominantly by different genotypes, and on the other hand, the use of PCA enables the defining elements of metabolic profiles to be distinguished. Furthermore, with metabolic analysis, response of metabolic networks to changes in single-gene loci is demonstrably complex, indicating how important it will be to have good methodologies in functional genomics that are capable of distinguishing cause from effect. Metabolite profiling is a valuable additional tool in the plant functional genomics repertoire and is worthy of wide application within and beyond the plant kingdom.

Experimental protocol

Arabidopsis plants were grown on GS 90 standard soil in growth chambers in a 16 h light / 8 h dark photoperiod, changing from 60% humidity and 20°C during the day to 75% humidity and 18°C at night. Light intensity was fixed to 120 µmol/m²/s. After approximately 8 h of the photoperiod, 300 mg fresh weight rosette leaves were harvested randomly from trays that had alternate lines of pots containing wild-type and transgenic plants ($n = 43$ (Col-2 WT), 45 (*dgd1*), 35 (C24 WT), and 28 (*sdd1-1*)). Extraction and fractionation was performed as reported recently[3]. Lipids were transmethylated by adding 900 µl chloroform and 1 ml methanol including 3% (vol/vol) sulfuric acid at 100°C for 4 h. Sulfuric acid was removed using three 4 ml portions of water. The lipophilic phase was dried over anhydrous sodium sulfate and carefully concentrated to about 80 µl. Before analysis, 20 µl of pyridine plus 20 µl of N-methyl-N-trimethylsilyl-trifluoroacetamide were added. $^{13}C_{12}$-Sucrose, $^{13}C_6$-glucose, d_8-glycerol, d_4-ethanolamine, d_6-ethylene glycol, d_3-aspartate, $^{13}C_5$-glutamate, d_4-alanine, d_8-valine, d_3-leucine, and d_5-benzoic acid were obtained from Campro Scientific (Emmerich, Germany) and used for exact quantification. GC/MS was performed using a GC 8000/Voyager mass spectrometer system (ThermoQuest, Manchester, UK). Peak finding and quantification of selective ion traces was accomplished using the instrument's MassLab FindTarget software. PCA and HCA pattern recognition was performed using the Pirouette software (Infometrix, Woodinville, WA) with $\log_{10}$ data transformation and mean-center preprocessing. Principal component analysis was performed with cross-validation. Hierarchical component analysis was performed using Euclidean distances with complete linkages.

Acknowledgments

This project was funded by the Max-Planck-Society. We thank Frank Kose, Una Griebel, and Antje Feller for their support in carrying out laboratory and computer work, Urte Schlüter for providing C24 WT and sdd1-1 mutant plants, and Megan McKenzie for revising the manuscript.

1. Somerville, C. & Somerville, S. Plant functional genomics. *Science* **285**, 380–383 (1999).
2. Baldwin, D., Crane, V. & Rice, D. A comparison of gel-based, nylon filter and microarray techniques to detect differential RNA expression in plants. *Curr. Opin. Plant Biol.* **2**, 96–103 (1999).
3. Santoni, V. *et al.* Use of a proteome strategy for tagging proteins present at the plasma membrane. *Plant J.* **16**, 633–641 (1998).
4. Trethewey, R.N., Krotzky, A.J. & Willmitzer, L. Metabolic profiling: a Rosetta stone for genomics? *Curr. Opin. Plant Biol.* **2**, 83–85 (1999).
5. Katona, Z.F., Sass, P. & Molnár-Perl, I. Simultaneous determination of sugars, sugar alcohols, acids and amino acids in apricots by gas chromatography–mass spectrometry. *J. Chromatogr. A* **847**, 91–102 (1999).
6. Adams, M.A., Chen, Z.L., Landman, P. & Colmer, T.D. Simultaneous determination by capillary gas chromatography of organic acids, sugars, and sugar alcohols in plant tissue extracts as their trimethylsilyl derivatives. *Anal. Biochem.* **266**, 77–84 (1999).
7. Duez, P., Kumps, A. & Mardens, Y. GC–MS profiling of urinary organic acids evaluated as a quantitative method. *Clin. Chem.* **42**, 1609–1615 (1996).
8. Fiehn, O., Kopka, J., Trethewey, R.N. & Willmitzer, L. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal. Chem.* **72**, 3573–3580 (2000).
9. McLafferty, F.W., Stauffer, D.A., Loh, S.Y. & Wesdemiotis, C. Unknown identification using reference mass spectra. Quality evaluation of databases. *J. Am. Soc. Mass Spectrom.* **10**, 1229–1240 (1999).
10. *Metabolite profiling* (Max Planck Institute of Molecular Plant Physiology). http://www.mpimp-golm.mpg.de/fiehn/index-e.html.
11. Gygi, S.P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994–999 (1999).
12. Vingron, M. & Hoheisel, J. Computation aspects of expression data. *J. Mol. Med.* **77**, 3–7 (1999).
13. Dörmann, P., Hoffmann-Benning, S., Balbo, I. & Benning, C. Isolation and characterization of an *Arabidopsis* mutant deficient in the thylakoid lipid digalactosyl diacylglycerol. *Plant Cell* **7**, 1801–1810 (1995).
14. Härtel, H., Lokstein, H., Dörmann, P., Trethewey, R.N. & Benning, C. Photosynthetic light utilization and xantophyll cycle activity in the galactolipid deficient *dgd1* mutant of *Arabidopsis thaliana*. *Plant Physiol. Biochem.* **36**, 407–417 (1998).
15. Dörmann, P., Balbo, I. & Benning, C. *Arabidopsis* galactolipid biosynthesis and lipid trafficking mediated by DGD1. *Science* **284**, 2181–2184 (1999).
16. Berger, D. & Altmann, T. A subtilisin-like serine protease involved in the regulation of stomatal density and distribution in *Arabidopsis thaliana*. *Gene Dev.* **14**, 1119–1131 (2000).
17. Mead R., Curnow R.N. & Hasted A.M. (eds). *Statistical methods in agriculture and experimental biology*, Edn. 2. (Chapman & Hall, London, 1993).
18. Murakami, Y., Tsuyama, M., Kobayashi, Y., Kodama, H. & Iba, K. Trienoic fatty acid and plant tolerance of high temperature. *Science* **287**, 476–479 (2000).
19. *GenomeNet database*. (Institute for Chemical Research, Kyoto University, Japan). http://www.genome.ad.jp/
20. *What is There? Interactive metabolic reconstruction on the WEB*. (Argonne Computational Biology Group, Chicago, IL). http://wit.mcs.anl.gov/WIT2/
21. Zweiger, G. Knowledge discovery in gene-expression-microarray data: mining the information output of the genome. *Trends Biotechnol.* **17**, 429–436 (1999).
22. Jurs, P.C. Pattern recognition used to investigate multivariate data in analytical chemistry. *Science* **232**, 1219 (1986).

The Plant Journal (2000) **23**(1), 131–142

TECHNICAL ADVANCE

Simultaneous analysis of metabolites in potato tuber by gas chromatography–mass spectrometry

Ute Roessner*, Cornelia Wagner, Joachim Kopka†, Richard N. Trethewey† and Lothar Willmitzer
Max-Planck-Institut für Molekulare Pflanzenphysiologie, Am Mühlenberg 1, 14476 Golm, Germany

Received 5 January 2000; revised 3 March 2000; accepted 1 April 2000.
*For correspondence (fax +49 331 567 8201; e-mail roessner@mpimp-golm.mpg.de).
†Present address: Metanomics GmbH and Co KGaA, Tegeler Weg 33, 10589 Berlin, Germany.

Summary

A new method is presented in which gas chromatography coupled to mass spectrometry (GC–MS) allows the quantitative and qualitative detection of more than 150 compounds within a potato tuber, in a highly sensitive and specific manner. In contrast to other methods developed for metabolite analysis in plant systems, this method represents an unbiased and open approach that allows the detection of unexpected changes in metabolite levels. Although the method represents a compromise for a wide range of metabolites in terms of extraction, chemical modification and GC–MS analysis, for 25 metabolites analysed in detail the recoveries were found to be within the generally accepted range of 70–140%. Further, the reproducibility of the method was high: the error occurring in the analysis procedures was found to be less than 6% for 30 out of 33 compounds tested. Biological variability exceeded the systematic error of the analysis by a factor of up to 10. The method is also suited for upscaling, potentially allowing the simultaneous analysis of a large number of samples. As a first example this method has been applied to soil- and *in vitro*-grown tubers. Due to the simultaneous analysis of a wide range of metabolites it was immediately apparent that these systems differ significantly in their metabolism. Furthermore, the parallel insight into many pathways allows some conclusions to be drawn about the underlying physiological differences between both tuber systems. As a second example, transgenic lines modified in sucrose catabolism or starch synthesis were analysed. This example illustrates the power of an unbiased approach to detecting unexpected changes in transgenic lines.

Keywords: GC-MS, metabolite, methoximation, potato tuber, trimethylsilylation.

Introduction

A central interest of our laboratory relates to the understanding of carbohydrate metabolism using the potato as a model plant. To this end we have created numerous transgenic plants characterized by either antisense inhibition or ectopic overexpression of proteins mediating pathways of carbohydrate metabolism. In order to understand better the complexity of events occurring at the metabolite level, it is also necessary to analyse various stages of development, looking at metabolites in many different pathways.

We therefore decided to develop a method that would allow the rapid, highly sensitive and quantitative simultaneous analysis of various metabolites indicative for carbon and nitrogen metabolism, such as sugars, sugar alcohols, dimeric and trimeric saccharides, amines, amino acids and organic acids. In addition, this method should be open to the range of metabolites analysed, and should be suitable for the development of automatic compound identification and quantification.

Here we present an approach essentially based on gas chromatography linked to mass spectrometry (GC–MS), which goes a long way towards meeting the above criteria. When applied to the polar phase of potato tuber extracts, this method allows the detection of a total of 150 compounds in a single run of 1 h duration. Among these 150 compounds we identified 77 with respect to their

chemical nature. The reproducibility of this method was high: the cumulative standard deviation was 6% or less for 30 out of 33 compounds tested with respect to the extraction, chemical modification and GC–MS analysis steps. For a wide range of compounds the biological variability was significantly in excess of the variability introduced by the method.

As a first example we applied this analytical approach to two types of potato tuber: soil-grown, and artificially induced in tissue culture. Whereas earlier data have suggested that these two systems are highly comparable, the broad metabolic profiling method described here reveals major and significant differences between the two systems.

We subsequently analysed a range of transgenic plants displaying a modified sucrose or starch metabolism. In addition to the insights provided by the profiles, the identification of several disaccharides including trehalose, maltose and isomaltose, and sugar alcohols including maltitol, specifically in individual transgenic lines illustrates the power of open metabolic profiling for detecting unexpected events.

Results and Discussion

Rationale for developing an analytical system based on GC–MS

Analysis of metabolites in plants is still often performed using either specific enzyme assays or chromatographic separations such as GC or HPLC, which take retention time/coelution as the main parameter for compound identification. Although powerful, these approaches suffer from at least two drawbacks: they represent non-open, biased approaches (the investigator will only find information about the metabolite that was experimentally targeted by the particular analysis protocol); and they are labour intensive because the detection systems used either provide information for only a single compound per assay (as is typical for enzyme-based metabolite assays) or, in the case of chromatographic separation coupled to non-specific detection, can only be applied to mixtures of low complexity, which in turn often necessitates clean-up steps.

Therefore approaches should be developed that combine high sensitivity and high specificity (such as that achieved by enzyme tests) with the potential to accommodate the analysis of highly complex mixtures of compounds. Mass spectrometry obviously meets these requirements as it combines high specificity, based upon compound specific fragmentation patterns, with high sensitivity. It has also been used extensively in plant sciences, although as a rule it has been applied for identifying a small number of specific compounds or for identification of compounds in extracts displaying a low complexity (e.g. Delarue *et al.*, 1998; Kamboj *et al.*, 1999; Yamaguchi *et al.*, 1990). In order to apply the selectivity of mass spectrometry to complex mixtures, a previous chromatographic separation is required. For this purpose either GC, liquid chromatography (LC) or capillary electrophoresis can be linked to mass spectrometric detection (e.g. Godber and Parsons, 1998; Katona *et al.*, 1999; Prinsen *et al.*, 1998; Tanaka *et al.*, 1998). In order to automatically identify a compound in a complex mixture the retention time is an essential parameter, as the analysis software needs to be able to limit the window within which it searches for a particular mass spectrum. Therefore reliability and reproducibility of retention time under given conditions is of crucial importance in the choice of separation method.

Comparison of the three methods mentioned above revealed that GC separation best fulfilled the criterion of high reproducibility in retention time for a set of given compounds. For this reason we decided to develop a method based GC separation and linked to a quadruple mass detector. However, GC–MS is not sufficient for a comprehensive analysis of plant metabolites as it is limited to those classes of compounds that are or can be made volatile, and thus can pass the GC separation under the conditions applied. There is undoubtedly scope for profiling methods to be developed for the more recent LC–MS technologies.

Extraction of polar metabolites from potato tubers followed by methoximation and silylation results in complex and reproducible GC–MS chromatograms of >150 compounds

Metabolites present in organisms such as higher plants are composed of multiple classes of chemical compounds. In order to be able to identify and quantify as many different metabolites as possible by the most reliable and least labour-intensive method, we tested a variety of extraction and chemical modification methods.

With respect to extraction, the addition of methanol to frozen plant material followed by a short heat treatment was found to give the most satisfying results. The combination of methanol and high (70°C) temperature is known to inactivate a majority of enzymes in several systems (Bligh and Dyer, 1959; Katona *et al.*, 1999), which is a necessary prerequisite to preventing changes in metabolite composition due to enzymatic conversions in the homogenate.

In order to make various classes of compounds volatile and thus accessible for analysis by GC, modification of polar functional groups is necessary. To find the most efficient derivatization reagents capable of working with a wide range of compound classes, we tested several

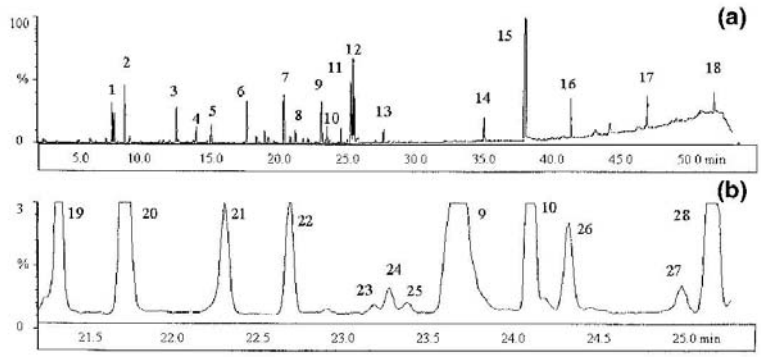


Figure 1. GC–MS total ion chromatogram of a tuber extract from *Solanum tuberosum* L. cv. Desirée, processed and analysed as described in Experimental procedures.
(a) Complete chromatogram, 4.0–50.0 min.
(b) Demonstration of sample complexity and analyte range by a representative expansion of the chromatogram in (a) for the region 21.5–25.0 min.
Peak identification: 1, glyceraldehyde MEOX1 TMS; 2, heptanoic acid TMS (time reference); 3, phosphoric acid TMS; 4, nonanoic acid TMS (time reference); 5, unknown substance; 6, malic acid TMS; 7, ribitol TMS (quantitative internal standard); 8, undecanoic acid TMS (time reference); 9, asparagine, *N,N,O*-TMS; 10, tridecanoic acid TMS (time reference); 11, glucose MEOX1 TMS; 12, citric acid TMS; 13, pentadecanoic acid TMS (time reference); 14, nonadecanoic acid TMS (time reference); 15, sucrose TMS; 16, tricosanoic acid TMS (time reference); 17, heptacosanoic acid TMS (time reference); 18, hentriacontanoic acid TMS (time reference); 19, glutamic acid, *N,O,O*-TMS; 20, pyroglutamic acid, *N,O*-TMS; 21, glutamine, *N,N,N,O*-TMS; 22, phenylalanine, *N,O*-TMS; 23, glucoheptonic acid TMS; 24, ribonic acid TMS; 25, unknown substance; 26, unknown substance; 27, mannitol TMS; 28, quinic acid TMS. Derivatives are per-trimethylsilylated unless otherwise indicated.

trimethylsilylation reagents: *N,O*-bis(trimethylsilyl)acetamide (BSA); trimethylsilylimidazole (TMSI); bis(trimethylsilyl)trifluoroacetamide (BSTFA); 1,1,1,3,3,3-hexamethyldisilazane (HMDS); and *N*-methyl-*N*-(trimetylsilyl)trifluoroacetamide (MSTFA) (Katona *et al.*, 1999). Among those reagents, MSTFA gave the best results with the broadest range of chemical compounds and produced the least by-products. Thus all subsequent silylation reactions were carried out using this reagent.

The carbonyl groups in sugars and sugar derivatives are another functional group requiring modification before GC–MS analysis. Methoxymation prevents ring formation by reducing sugars and stabilizes carbonyl moieties in the β-position, and has been described as a suitable approach (Schweer, 1982). We therefore studied different conditions for the methoximation reaction. Methoximation produces two different stereoisomers that were resolved by our chromatography. It is vital that the methoximation reaction proceeds to completion, otherwise a third peak representing the non-methoxyaminated substance will appear. For this reason, reaction times and temperatures of the methoximation procedure were optimized. The best results were obtained following incubation with methoximation reagent for 90 min at 30°C.

The chemically modified extracts were subsequently subject to GC–MS analysis, leading to peak resolution and signal responses that were acceptable for both qualitative identification and quantification purposes. A typical example of a GC–MS total ion chromatogram from the polar phase of a methanol of greenhouse-grown potato tubers is shown in Figure 1.

In Figure 1 a complex pattern of major, minor and trace peaks can be observed. If the amount of a specific compound exceeds the column capacity or the linear response range of the detector then a subsequent analysis should be performed with a higher sample dilution.

When two compounds coelute, differences in the respective mass spectra (leading to different ions specific for the compounds in question) can be used selectively to quantify both compounds. Sensitivity can be increased further by running the quadrupole mass detector in the single ion-monitoring mode. In this mode only a small mass range is analysed, and many more scans are performed per time unit resulting in an approximately 30-fold increase in sensitivity for a given *m/z* ratio.

77 compounds of known chemical structure can be identified in the polar fraction of potato tuber extracts

As shown in Figure 1 the total ion chromatogram displays the peak profile of a complex mixture. In order to identify the chemical nature of as many peaks as possible, two strategies were followed. First, the spectra of all identifiable peaks were compared with commercially available electron impact mass spectrum libraries such as NIST for Masslab (Fisons, Manchester, UK) or WILEY (Palisade

Table 1. Metabolites identified from a methanol tissue extract of potato tubers

Amino acids	Organic acids	Sugars	Sugar alcohols	Aromatic amines
β-alanine (248)	2-aminoadipic acid (260)	erythrose (205)	erythritol (307)	dopamine (426)
alanine (116)	ascorbic acid (332)	fructose (307)	glycerol (205)	noradrenaline (355)
arginine (348)	benzoic acid (179)	fructose-6-phosphate (315)	inositol (305)	normetanephrine (297)
asparagine (188)	citric acid (465)	fucose (160)	maltitol (361)	octopamine (426)
aspartic acid (232)	fumaric acid (245)	galactose (160)	mannitol (319)	tyramine (338)
cystathionine (278)	glucoheptonic acid (435)	glucose (160)	xylitol/arabitol (319)	
cysteine (220)	gluconic acid (333)	glucose-6-phosphate (387)		
GABA (304)	glutaric acid (261)	isomaltose (480)		
glutamic acid (246)	glyceric acid (189)	mannose (160)		
glutamine (156)	hydroxybenzoic acid (267)	maltose (361)		
glycine (174)	isocitric acid (273)	raffinose (437)		
histidine (154)	α-ketoglutaric acid (198)	ribose (307)		
homocysteine (234)	lactic acid (117)	sucrose (451)		
homoglutamine (170)	malic acid (245)	trehalose (191)		
homoserine (218)	oxalic acid (147)	xylose/arabinose (307)		
isoleucine (158)	oxaloacetic acid (333)			
leucine (158)	6-phospho-gluconic acid (387)			
L-hydroxyproline (230)	3-phospho-glycerate (299)			
lysine (317)	phosphoric acid (299)			
methionine (176)	quinic acid (345)			
ornithine (142)	ribonic acid (307)			
phenylalanine (218)	shikimic acid (372)			
proline (142)	succinic acid (247)			
serine (116)				
threonine (291)				
tryptophane (202)				
tyrosine (218)				
valine (144)				

Identification was performed by demonstration of cochromatography of a standard substance and by comparison of the mass spectrum to the standard (see Experimental procedures). The *m/z* value indicated after each compound is the specific ion used for quantifying each metabolite. GABA, gamma-amino butyric acid.

Cooperation, New York, USA). Secondly, GC–MS analysis was performed using several hundred standard compounds which we assumed to be present in detectable amounts in plant tissues, thus creating our own reference library containing both the retention index of these compounds (as determined under our conditions) and the corresponding mass spectrum.

When these approaches were applied to extracts from potato tubers of non-genetically modified reference plants and to various transgenic lines, we could identify a total of 77 compounds (Table 1). However, there were still a large number of peaks which were not found either in any of the commercial libraries or in the several hundred compounds that we tested directly. These peaks will require significant further efforts before a chemical name and structure can be assigned to them. The mass spectra of all standard compounds that were processed can be found at www.mpimp-golm.mpg.de/willmitzer/index-e.html.

As shown in Table 1, we were also able to detect various sugar phosphates. Although surprising at first glance due to the assumed low volatility of these substances, the mass spectrum clearly identified the phosphorylated compounds due to the presence of an M-15 fragment peak at *m/z* = 706 for both glucose-6-phosphate and fructose-6-phosphate. The appearance of free phosphate-TMS esters (*m/z* = 314) at the same retention time further indicates that intact sugar phosphates enter the ionization source.

Variability caused by extraction, chemical modification and analysis by GC–MS is small when compared to the biological variability within samples

An essential factor in assessing the quality of analytical processes is the reproducibility of results. In order to test our system with respect to this parameter, we divided the analytical process into three components that contribute to the observed variability. These were variability caused by: (i) the final analysis (chromatography, detection, stability of chemically modified samples); the sample preparation; the biological material.

To gain an insight into the reproducibility of the final analysis under authentic conditions, a single chemically modified sample composed of 23 representative standard substances aliquoted in separate measurement vials was measured every 2 h over a 30 h period (the processing time of a typical series of 30 samples).

Table 2. Reproducibility of derivatization reaction and GC–MS analysis using potato tuber extracts from plants expressing an invertase in the cytosol (Sonnewald *et al.*, 1997)

Derivative[a]	Response ratio[b]	SD	SD (%) ($n=20$)
alanine, N,O-TMS	0.72	0.0592	8.28
asparagine, N,N,O-TMS	2.78	0.1126	4.06
aspartic acid, N,O,O-TMS	0.87	0.0096	1.11
cysteine, N,O,S-TMS	0.01	0.0007	6.65
glutamic acid, N,O,O-TMS	1.07	0.0194	1.82
glutamine, N,N,O-TMS	1.61	0.0776	4.81
glycine, N,N,O-TMS	0.17	0.0061	3.52
isoleucine, N,O-TMS	0.28	0.0146	5.18
leucine, N,O-TMS	0.15	0.0093	6.12
lysine, N,N,N,O-TMS	0.03	0.0009	2.68
methionine, N,O-TMS	0.32	0.0075	2.36
phenylalanine, N,O-TMS	0.59	0.0051	0.86
proline, N,O-TMS	0.04	0.0027	7.31
pyroglutamic acid, N,O-TMS	1.64	0.0328	2.01
serine, N,O,O-TMS	0.51	0.0249	4.88
threonine, N,O,O-TMS	0.04	0.0017	4.76
tyrosine, N,N,O-TMS	0.76	0.0398	5.26
valine, N,O-TMS	0.73	0.0410	5.61
citric acid TMS	6.78	0.3896	5.74
fumaric acid TMS	0.02	0.0009	4.89
isocitric acid TMS	0.05	0.0017	3.39
malic acid TMS	0.49	0.0039	0.82
quinic acid TMS	0.08	0.0018	2.36
shikimic acid TMS	0.01	0.0003	2.37
succinic acid TMS	0.06	0.0017	3.01
fructose MEOX1 TMS	0.01	0.0003	3.32
glucose MEOX1 TMS	8.36	0.4605	5.51
glucose-6-P MEOX1 TMS	0.27	0.0127	4.68
mannose MEOX TMS	0.03	0.0010	3.01
sucrose TMS	0.05	0.0014	2.34
glycerol TMS	0.01	0.0019	14.10
inositol TMS	0.03	0.0006	2.04
mannitol TMS	0.39	0.0031	0.78

Tuber extracts from 6 representative developing tubers were pooled and divided into 20 aliquots. Each aliquot was derivatized and analysed as described in Experimental procedures. MEOX, methoxyaminated derivative; TMS, trimethylsilylated derivative; SD, standard deviation.
[a]Derivatives were per-trimethylsilylated if not otherwise indicated.
[b]Response ratios represent peak area ratios using ribitol as quantitative internal standard.

Table 3. Variability of individual potato tubers

Derivative[a]	Relative response ratio[b] (g^{-1} FW)	SD[a] ($n=9$)	%[a]
alanine, N,O-TMS	10.18	2.595	25.48
asparagine, N,N,O-TMS	112.04	26.693	23.82
aspartic acid, N,O,O-TMS	32.08	3.085	9.62
cysteine, N,O,S-TMS	0.39	0.102	25.98
glutamic acid, N,O,O-TMS	66.05	5.582	8.45
glutamine, N,N,O-TMS	179.92	40.421	22.47
glycine, N,N,O-TMS	4.64	0.758	16.33
isoleucine, N,O-TMS	11.03	1.262	11.44
leucine, N,O-TMS	2.87	0.372	12.94
lysine, N,N,N,O-TMS	8.84	1.072	12.12
methionine, N,O-TMS	18.75	2.925	15.60
phenylalanine, N,O-TMS	30.46	5.953	19.54
proline, N,O-TMS	1.81	0.252	13.95
pyroglutamic acid, N,O-TMS	94.19	20.277	21.53
serine, N,O,O-TMS	10.38	1.645	15.84
threonine, N,O,O-TMS	2.88	0.452	15.70
tyrosine, N,N,O-TMS	59.93	12.457	20.79
valine, N,O-TMS	25.09	2.520	10.04
citric acid TMS	553.42	56.165	10.15
fumaric acid TMS	0.20	0.031	15.17
isocitric acid TMS	3.48	0.464	13.33
malic acid TMS	11.02	4.181	37.94
quinic acid TMS	5.16	0.873	16.93
shikimic acid TMS	0.42	0.055	13.22
succinic acid TMS	0.25	0.074	29.97
fructose MEOX1 TMS	6.01	3.374	56.14
glucose MEOX1 TMS	93.48	53.754	57.50
glucose-6-P MEOX1 TMS	1.86	0.231	12.40
sucrose TMS	168.36	34.699	20.61
inositol TMS	9.44	0.913	9.67
mannitol TMS	6.28	0.944	15.05
mannose MEOX TMS	0.28	0.068	24.28

Samples were obtained from nine wild-type potato tubers of a single harvest. Extraction, derivatization, and analysis were performed as described in Experimental procedures.
[a]Derivatives were per-trimethylsilylated if not otherwise indicated.
[b]Response ratio was normalized with respect to the fresh weight of the tuber sample.

In this experiment most compounds yielded stable results. In general a loss of only 2% of the initial signal was observed by the end of the 30 h analysis period (data not shown). The only exceptions were glutamine, tryptophane and glycerol, where losses of 24, 10 and 9%, respectively, were observed. We therefore conclude that under typical conditions the reliability of the method is very high.

Before the samples are injected into the GC–MS they undergo a complex series of extraction and chemical modification steps. To see how these steps influence the variability of the results obtained, a single tuber extract was divided into 20 aliquots, chemically processed and finally analysed by GC–MS. Analysis of the data shown in Table 2 demonstrate that the standard deviation introduced by these steps is, as a rule, below 6% of the mean. This was the case for 29 of the 33 metabolites tested in this experiment.

As both of these analyses demonstrate the robustness and high reproducibility of the analytical method *per se*, we decided to test the biological variability of samples. Tuber slices from nine individual wild-type plants grown side-by-side under identical conditions were analysed. Under ideal growth conditions the metabolites display

Table 4. Quantitative determination of metabolite concentrations in developing potato tubers

Metabolite	Concentration (μmol g^{-1} FW)	SE ($n=6$)	%
β-alanine	0.15	0.01	6.6
alanine	1.68	0.34	20.2
ascorbic acid	0.53	0.14	26.4
asparagine	5.62	2.01	35.7
aspartic acid	1.27	0.11	8.7
cysteine	0.41	0.01	2.4
glutamine	1.08	0.26	24.1
glycine	0.23	0.03	13.0
isoleucine	0.94	0.16	17.0
leucine	0.43	0.12	27.9
lysine	1.06	0.16	15.1
methionine	0.65	0.08	12.3
phenylalanine	0.59	0.09	15.2
proline	0.18	0.01	5.5
serine	1.34	0.15	11.2
valine	2.51	0.29	11.5
fumaric acid	0.20	0.01	2.8
glyceric acid	0.15	0.01	3.3
citric acid	18.86	3.27	17.3
isocitric acid	0.17	0.04	24.1
malic acid	5.39	0.73	13.5
oxalic acid	1.00	0.19	19.5
quinic acid	15.67	2.36	15.1
shikimic acid	0.37	0.02	5.8
succinic acid	0.97	0.10	10.6
fructose	0.01	0.00	25.0
galactose	0.01	0.00	23.5
glucose	23.84	4.09	17.1
glucose-6-phosphate	0.21	0.00	1.8
mannose	0.14	0.00	4.7
sucrose	25.91	3.29	12.7
inositol	0.06	0.01	11.7
mannitol	0.06	0.00	6.1

Single tuber samples of six plants were measured. Developing tubers were harvested during the spring season from 10-week-old greenhouse plants grown in 2 l pots. SE, Standard error.

standard deviations below 20% of the mean (Table 3) with only two exceptions, glucose and fructose, which are well known to demonstrate a higher variability in potato tubers (Merlo *et al.*, 1993; Veramendi *et al.*, 1999a; Veramendi *et al.*, 1999b). In most cases biological variability exceeded the experimental error by at least a factor of two (observed range 1.5- to 10-fold).

We therefore conclude that with the GC–MS technology the variability in results is essentially due to the variability within the biological samples themselves.

GC–MS allows both absolute and relative determination of the compounds detected

In order to extend the evaluation of the suitability of this GC–MS approach for metabolic profiling, we quantified representative metabolites in potato tubers. We established calibration curves for 33 compounds, members of five chemical classes: amino acids, organic acids, sugars, saccharides and sugar alcohols. A linear relationship covering the normal concentration range present in plant tissues was observed (data not shown).

These metabolites were then quantified in developing potato tubers (Table 4); the resulting absolute levels with respect to fresh weight were found to be in the same range as previously reported by both our and other groups using enzymatically linked photometric assays or HPLC analysis of extracts (Burrell *et al.*, 1994; Geigenberger *et al.*, 1998; Sweetlove *et al.*, 1996; Trethewey *et al.*, 1998). Thus we judged this technology valid. The only major exception observed was with respect to the citric acid content. Citric acid levels were approximately 10 times higher using the GC–MS approach in comparison with previous studies using spectrophotometric-based methods. We were not able to determine the reason for this discrepancy.

In addition to the quantitative measurements, we performed a series of recovery experiments for 25 metabolites drawn as representatives of these five compound classes, and found that the calculated recoveries were within the generally accepted range for analysis work at 70–140% (see Experimental procedures). Taking together the results of the analysis of variability, the data from the recovery experiments, and the comparability of the absolute values determined with previous studies, we conclude that the GC–MS-based approach is valid for the study of metabolism in potato tubers.

In many cases absolute concentrations are not of prime importance. For the analysis of a special environmental or developmental situation, or for the comparison of a specific genotype with standard tissue samples, relative data are sufficient. The current profiling techniques applied in RNA expression analysis provide only relative data. The determination of relative values can be easily achieved by the GC–MS method as well. For this purpose, within each chromatogram the peak areas derived from specific ion traces indicative of each analysed compound are normalized by the peak area derived from an internal standard, such as ribitol, present within the same chromatogram, resulting in response ratios for all compounds analysed (see Experimental procedures). The response ratios are subsequently converted to relative response ratios through division by the fresh weight of each sample, thus achieving further normalization. These relative response ratios can be directly compared between different samples without knowledge of absolute compound concentrations. The specific ion masses used for the respective analyte for quantification are given in Table 1.

Determination of both relative values and absolute concentrations have advantages and disadvantages. However, for comparative purposes between two sample

types we routinely describe changes by calculating the quotients of the various relative response ratios for each compound.

Automization of the chromatogram analysis

It is obvious from the preceding descriptions that evaluation by eye of each single chromatogram would be an impossible task and would represent a major stumbling block for the efficient use of this technology. We therefore decided to use a chromatography data analysis algorithm that allows multiparallel and automatic identification and quantification of compounds present in a GC–MS chromatogram. This basic algorithm, which is part of the MASSLAB software distributed by ThermoQuest (Manchester, UK), allows the use of two strategies for identification of a compound and quantification of a set of given compounds. For identification, the first parameter used by the software is the retention time of the compound relative to a standard compound. Initially we used just one compound for determination of relative retention times; however, this proved insufficient as non-linear shifts of retention times occur in the elution profiles during the lifetime of a GC column. We therefore decided to spike every sample before injection with several compounds that were absent from the fraction under analysis. For the work presented here we used fatty acids (Figure 1); the use of fatty acids has been described as an alternative approach to using *n*-alkanes for determining the Kovats index (Castello, 1999; Gonzalez and Nardillo, 1999). Subsequently the relative retention time of each compound is interpolated with respect to its two nearest neighbours and taken as the first identification criterion. As a second subsequent criterion we used the full electrone impact ionization (EI) mass spectrum for the compounds in question. A match factor was automatically calculated which provides an indication of the reliability of assignment. In other words, a confidentiality index is provided which alerts the investigator when manual inspection of the compound assignment is necessary.

For quantification purposes a single specific ion known to be selective and sensitive for the compound in question was taken, and the response ratio calculated as described above. For the analysis presented here ribitol was used as the quantitative internal standard (see Experimental procedures).

Applying this method to potato tuber extracts, we can currently identify and automatically quantify 60 of the 77 known compounds with a very high fidelity, i.e. less than 0.1% manual inspection. The remaining 18 compounds can be quantified, but this process requires at least 5% reassessments of the original chromatogram by a trained user. It is to be expected that with the development of improved software and GC–MS hardware the number of compounds requiring supplementary manual assessment will decrease.

Metabolic profiling using GC–MS reveals major differences between soil- and in vitro*-grown tubers*

When analysing the growth and development of plant systems, *in vitro* culture is an interesting technology as it allows the manipulation of environmental conditions in a defined manner. With respect to potato tubers, numerous reports including work from our own group have described the suitability of *in vitro*-grown tubers for the analysis of gene expression or changes in metabolite composition. Morphological, molecular and biochemical data appear to be in agreement with the assumption that *in vitro*-grown tubers represent a faithful phenocopy of soil-grown potato tubers (Debon *et al.*, 1998; Desire *et al.*, 1995; Veramendi *et al.*, 1999a; Visser *et al.*, 1994). In view of the compelling evidence for the equivalence of these systems, we decided to apply the metabolic profiling approach in order to assess a broader range of compounds using this unbiased approach.

Data describing the ratios of many metabolites in both systems are summarized in Figure 2, which shows that *in vitro*- and soil-grown potato tubers are not as similar as previously thought. Major differences were found in the group of amino acids, most notably in glutamic acid and other amino acids derived from α-ketoglutaric acid such as glutamine, proline and arginine. In addition, large increases were observed in the levels of amino acids derived from oxaloacetic acid, such as asparagine and lysine, within *in vitro*-grown tubers. However, tyrosine, glycine, alanine, β-alanine and phenylalanine decreased in *in vitro*-grown when compared to soil-grown tubers. In general, microtubers were found to contain a much higher amount of amino acids in comparison to soil-grown tubers. A high anaplerotic demand for carbon may explain the decreases observed in citric acid-cycle intermediates such as citric acid, malic acid, succinic acid and isocitric acid. Alternatively, many of these changes could be a consequence of the higher amount of nitrate supplied to the microtubers. As the *in vitro*-grown tubers are probably not limited in carbohydrate supply due to the large amount of sugars present in the tuber-inducing medium, the rate of nitrate assimilation may be markedly increased in comparison with soil-grown tubers (Morcuende *et al.*, 1998).

Another important difference between both types of tubers is the higher amount of compounds indicative of osmotic stress found in the *in vitro* tubers (Hare *et al.*, 1998; Hare *et al.*, 1999; Karakas *et al.*, 1997). For exam-

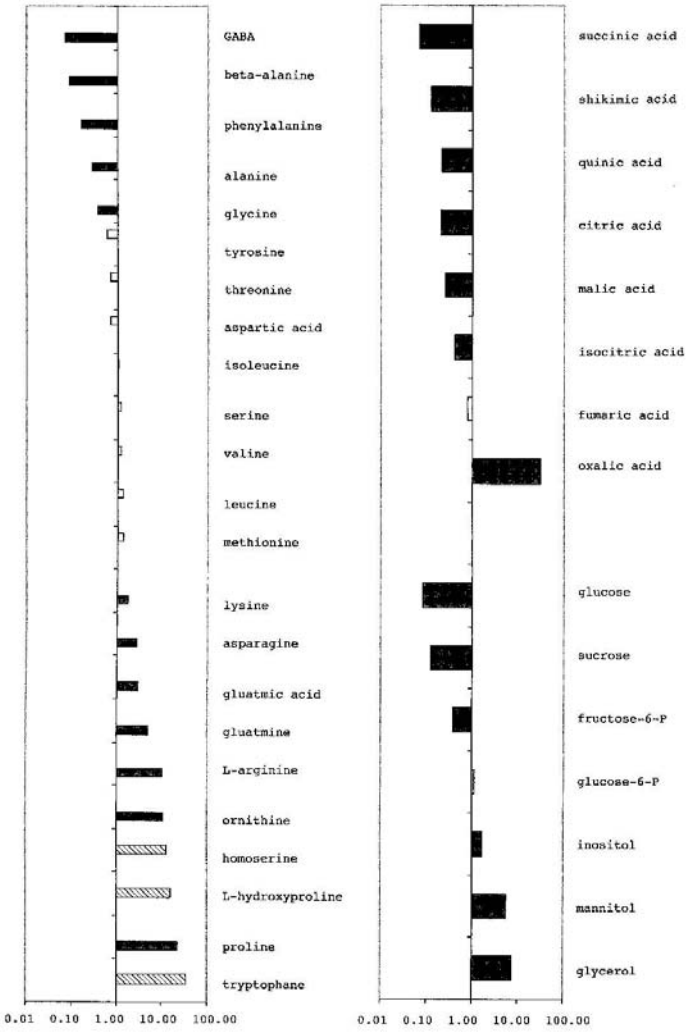


Figure 2. Comparison of metabolite levels in developing soil-grown tubers and *in vitro* tubers.
The quotient of mean relative response ratio from *in vitro* tuber samples ($n=6$) and mean relative response ratio of soil-grown tuber samples ($n=6$) are plotted using a logarithmic scale. Values <1 represent a decrease in metabolite levels in the *in vitro* tuber compared to soil-grown tuber; values >1 represent an increase. The significance of the changes was evaluated by a *t*-test; black bars indicate statistically significant differences between the systems ($P<0.05$), white bars show non-significant differences, grey bars indicate metabolites which were detected in the *in vitro* tubers but were below the detection limit for the soil-grown tuber samples. In these cases the numerical value of the detection limit of the respective compound within the soil-grown tuber samples was used to calculate an estimated, representative quotient.

ple, glycerol, mannitol, inositol, and proline are significantly increased. Finally, some of the changes seen in the amino acid pools, such as the increase in asparagine, glutamine, and glutamic acid, have been observed in water-stressed leaves of tomato (Bauer *et al.*, 1997), sugar beet (Gzik, 1996) and mulberry

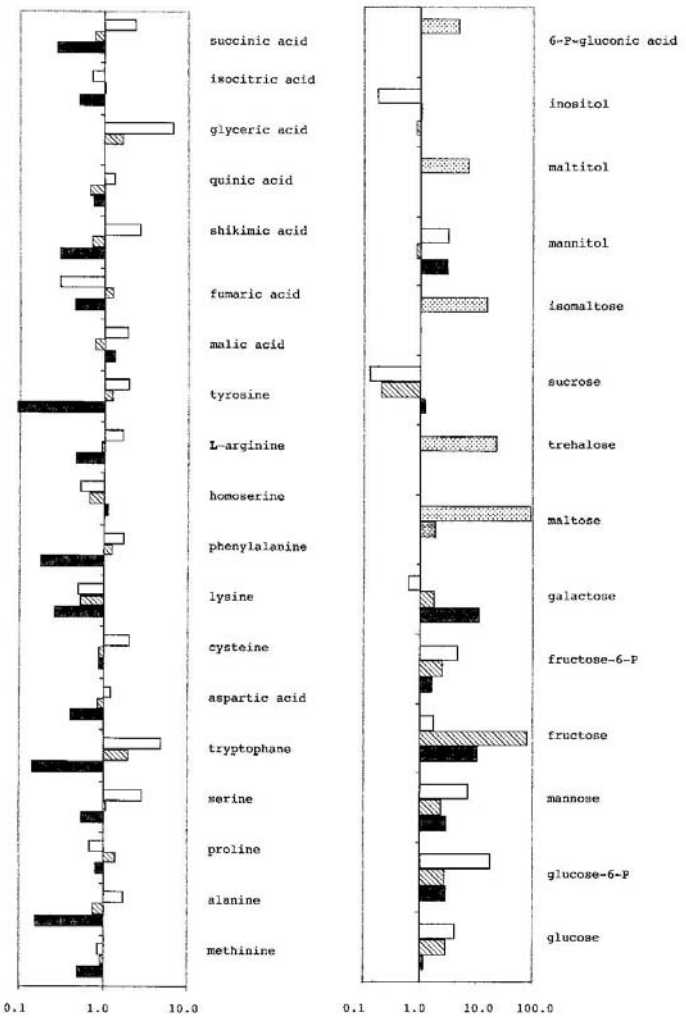


Figure 3. Comparison of metabolite levels in wild-type developing potato tubers with those in tubers of transgenic potato plants.
Transgenic plants exhibiting antisense repression of ADP-glucose pyrophosphorylase (AGP93, black bars), or overexpressing a yeast invertase in the apoplast (U-IN1–33, grey bars) or in the cytosol (U-IN2–30, white bars). The quotients of the mean relative response ratio from transgenic tuber samples ($n=6$) and from wild-type tuber samples ($n=6$) are plotted using a logarithmic scale, as described for Figure 2. Only changes in metabolite levels that were evaluated to be significantly different from the wild type are shown ($P<0.05$). Dotted bars indicate metabolites which were detected in the transgenic tubers but were below the detection limit in wild-type tubers. In these cases the numerical value of the detection limit of the respective compound in the soil-grown tuber samples was used in order to estimate a representative quotient. Metabolites that did not show significant changes between the four genotypes were leucine, isoleucine, beta-alanine, ornithine, valine, asparagine, glycine, glutamine, threonine, glutamic acid and GABA.

(Ramanjulu and Sudhakar, 1997). The reason for the possible water stress could be the high amount of sucrose present in the tuber-inducing medium.

In conclusion, the application of a broad profiling method to two systems previously thought to be very comparable reveals that caution must be exercised when

interpreting biological events by the analysis of a restricted number of parameters.

Metabolic profiling of several transgenic lines modified in either sugar or starch metabolism reveals unexpected changes in disaccharides and sugar alcohols

In order to explore further the power of metabolic profiling for detecting unexpected changes, tubers from three transgenic and one wild-type line were subjected to GC–MS analysis, as described above. The transgenic lines analysed were either altered in sucrose metabolism (expression of a yeast invertase in either the cytosol or apoplast (Sonnewald *et al.*, 1997; Trethewey *et al.*, 1998), or inhibited in starch biosynthesis following antisense RNA-mediated repression of the activity of ADP-glucose pyrophosphorylase (Müller-Röber *et al.*, 1992).

GC–MS analysis revealed that amongst the metabolites analysed a large fraction (33) were significantly changed in at least one of the transgenic lines. A comparison of the data obtained for the three transgenic lines using GC–MS profiling with the values previously published from HPLC analysis for amino acids and organic acids (Trethewey *et al.*, 1998; Trethewey *et al.*, 1999) shows that the profiling data are in broad agreement with the previously determined values (Figure 3). Our general experience suggests that the differences observed in the extent of quantitative changes are due to the normal differences between the growing conditions of different batches of tubers.

All the scientific conclusions previously drawn about the redistribution of metabolism in the transgenic lines can be drawn from this single profiling analysis. In particular, the increase in glycolysis, amino acids and organic acids in the line expressing the yeast invertase in the cytosol can be clearly deducted from the data in Figure 3. With respect to the transgenic line expressing a yeast invertase in the apoplast (U-IN1), it is apparent from Figure 3 that this line does not have an elevated respiratory flux.

With respect to the line expressing the yeast invertase in the cytosol, the appearance of 6-phosphogluconic acid is a new result and can be taken as a strong indication of a significantly increased level of the oxidative pentose phosphate pathway, which might be expected given the large increase in glucose-6-phosphate in this line.

A most remarkable and unexpected finding concerns the presence of dramatically increased levels of disaccharides such as maltose, isomaltose and trehalose in the line expressing the cytosolic invertase. This result has been found in three independent lines (data not shown) Specifically, the presence of trehalose is an exciting observation. In addition to confirming and extending previous observations on the capacity of higher plants to synthesize trehalose under specific conditions (Goddijn and Smeekens, 1998; Müller *et al.*, 1999), this finding is of relevance as trehalose has been discussed extensively in the literature as a possible signal for the metabolic status of a given cell. The fact that trehalose appears only in those cells where an increased rate of glycolysis and respiration has been observed makes this a fascinating and stimulating observation. Further comparisons with other lines displaying a change in the rate of glycolysis and respiration must reveal whether or not this exciting but nevertheless simple interpretation is true.

The transgenic lines have been analysed in significant detail in earlier publications by applying conventional techniques such as enzyme assays and HPLC. Nevertheless, the occurrence of trehalose and other disaccharides has not been noted, mainly due to the fact that the methods used previously were not capable of determining these compounds. The unbiased approach for metabolic profiling presented here gives a clear advantage in providing the opportunity to find unexpected events, as exemplified by the identification of trehalose, and thus may provide novel insights into metabolic networks.

Conclusions

Using GC–MS as a system for separation and detection of metabolites, we were able to develop a simple procedure allowing the simultaneous analysis of a large group of compounds representing sugars, sugar alcohols, amino acids, organic acids and some special compounds in plant extracts. This method allows an unbiased study of a range of metabolic pathways with only minimal effort. As this technology utilizes MS, it also combines high sensitivity with high specificity and, as shown here, high reproducibility.

We are presently in the process of applying this technique to a number of other plant species. Depending upon the species and the tissue, between 50 and 400 different compounds can be detected when the analysis is applied to hydrophilic and lipophilic classes of compounds. As this system has the potential to be fully automated, we believe that in the future it will represent a major tool for characterizing the metabolic status of a plant with respect to environmental, developmental or genetic factors.

Experimental procedures

Plant material

Potato plants, *Solanum tuberosum* L. cv. Desirée, were supplied by Saatzucht Lange (Bad Schwartau, Germany). Plants were grown in a greenhouse at 22°C under a 16 h light/8 h dark regime with supplementary artificial light under a minimum of 250 µmol photons $m^2 sec^{-1}$. In this paper the term 'developing tubers' refers to tubers above 10 g that were harvested from 10-week-old plants.

Microtubers were generated *in vitro* as described by Veramendi *et al.* (1999a). Single-node explants were cultured in MS media

(Murashige and Skoog, 1962) containing 6% (w/v) sucrose and 11.6 µM kinetin. Sets of 15 explants were kept in the dark in glass jars containing 50 ml MS medium. Microtubers were harvested after 4 weeks at 20°C.

Chemicals

All chemicals and pure standard substances were purchased from either Sigma–Aldrich Chemie GmbH (Deisenhofen, Germany) or Merck KGaA (Darmstadt, Germany).

Extraction and derivatization of potato tuber metabolites for GC–MS analysis

Tuber slices or whole microtubers were immediately frozen in liquid nitrogen and stored at −70°C until further analysis. A polar metabolite fraction was obtained from either approximately 50 mg microtuber or 100 mg tuber slices by Ultra Thurax T25 (IKA Labortechnik, Staufen, Germany), homogenization in 1400 µl 100% methanol with 50 µl internal standard (2 mg ribitol ml^{-1} water). The mixture was extracted for 15 min at 70°C. The extract was vigorously mixed with 1 vol water and subsequently centrifuged at 2200 ***g***. Aliquots of the methanol/water supernatant (1.00 or 0.25 ml) were dried *in vacuo* for 6–16 h. The dried residue was redissolved and derivatized for 90 min at 30°C (in 80 µl of 20 mg ml^{-1} methoxyamine hydrochloride in pyridine) followed by a 30 min treatment at 37°C (with 80 µl MSTFA). 40 µl of a retention time standard mixture was added prior to trimethylsilylation. This retention time standard mixture contained 3.7% (w) heptanoic acid, 3.7% (w) nonanoic acid, 3.7% (w) undecanoic acid, 3.7% (w) tridecanoic acid, 3.7% (w) pentadecanoic acid, 7.4% (w) nonadecanoic acid, 7.4% (w) tricosanoic acid, 22.2% (w) heptacosanoic acid and 55.5% (w) hentriacontanoic acid dissolved in tetrahydrofuran at 10 mg ml^{-1} total concentration.

Standard substances for peak identification were dissolved in water at 10 mg ml^{-1}. A 5 µl volume of standard solution was dried *in vacuo* and derivatized with 50 µl of 20 mg ml^{-1} methoxyamine hydrochloride in pyridine and 50 µl MSTFA, as described above.

To establish the efficiency of the extraction procedure, the recovery of various standard metabolites was determined by the addition of authentic metabolite standards to the tissue sample at the start of the extraction procedure. Standard substances were added in threefold excess of the determined endogenous concentrations. Estimates of recovery were 138% for alanine, 100% for aspartic acid, 107% for glycine, 107% for isoleucine, 99% for leucine, 92% for lysine, 106% for phenylalanine, 90% for proline, 118% for valine, 127% for serine, 110% for fructose, 78% for fructose-6-phosphate, 101% for galactose, 129% for glucose, 78% for glucose-6-phosphate, 115% for maltose, 100% for sucrose, 101% for inositol, 108% for mannitol, 139% for glyceric acid, 117% for fumaric acid, 98% for isocitric acid, 118% for malic acid, 99% for shikimic acid and 73% for succinic acid.

GC–MS analysis

Sample volumes of 1 µl were injected with a split ratio of 25:1 using a hot-needle technique. The GC–MS system consisted of an AS 2000 autosampler, a GC 8000 gas chromatograph and a Voyager quadrupole mass spectrometer (ThermoQuest, Manchester, UK). The mass spectrometer was tuned according to the manufacturer's recommendations using tris-(perfluorobutyl)-amine (CF43). Gas chromatography was performed on a 30 m SPB-50 column with 0.25 mm inner diameter and 0.25 µm film thickness (Supelco, Bellfonte, CA, USA). Injection temperature was 230°C, the interface set to 250°C and the ion source adjusted to 200°C. The carrier gas used was helium set at a constant flow rate of 1 ml min^{-1}. The temperature program was 5 min isothermal heating at 70°C, followed by a 5°C min^{-1} oven temperature ramp to 310°C and a final 1 min heating at 310°C. The system was then temperature equilibrated for 6 min at 70°C prior to injection of the next sample. Mass spectra were recorded at two scans per sec with an *m/z* 50–600 scanning range. The chromatograms and mass spectra were evaluated using the MASSLAB program (ThermoQuest, Manchester, UK). A retention time and mass spectral library for automatic peak quantification of metabolite derivatives was implemented within the MASSLAB method format.

Statistical analysis

The *t*-tests were performed using the algorithm incorporated into Microsoft EXCEL (Microsoft Corporation, Seattle, WA, USA). The word significant is used in the text when the change in question has been confirmed to be statistically significant ($P < 0.05$) with the *t*-test.

Acknowledgements

We would like to thank the gardeners for taking excellent care of the greenhouse plants. Our special thanks are assigned to Dr Alisdair Fernie for proofreading the manuscript and his helpful discussions. We would also like to thank Kristina Zubow for her patient and tireless efforts in creating the mass spectral library.

References

Bauer, D., Biehler, K., Fock, H., Carrayol, E., Hirel, B., Migge, A. and Becker, T.W. (1997) A role for cytosolic glutamine synthetase in the remobilization of leaf nitrogen during water stress in tomato. *Physiol. Plant.* **99**, 241–248.

Bligh, E.G. and Dyer, W.J. (1959) A rapid method of total lipid extraction and purification. *Can. J. Biochem. Physiol.* **31**, 911–917.

Burrell, M.M., Mooney, P.J., Blundy, M., Carter, D., Wilson, F., Green, J., Blundy, . K.S. and ap Rees, T. (1994) Genetic manipulation of 6-phosphofructokinase in potato tubers. *Planta*, **194**, 95–101.

Castello, G. (1999) Retention index systems: alternatives to the *n*-alkanes as calibration standards. *J. Chromatogr. A.* **842**, 51–64.

Debon, S.J.J., Tester, R.F., Millam, S. and Davies, H.V. (1998) Effect of temperature on the synthesis, composition and physical properties of potato microtuber starch. *J. Sci. Food Agric.* **76**, 599–607.

Delarue, M., Prinsen, E., van Onckelen, H., Caboche, M. and Bellini, C. (1998) *Sur2* mutations of *Arabidopsis thaliana* define a new locus involved in the control of auxine homeostasis. *Plant J.* **14**, 603–611.

Desire, S., Couillerot, J.P., Hilbert, J.L. and Vasseur, J. (1995) Protein changes in *Solanum tuberosum* during *in vitro* tuberization of nodal cuttings. *Plant Physiol. Biochem.* **33**, 303–310.

Geigenberger, P., Hajirezaei, M., Geiger, M., Deiting, U., Sonnewald, U. and Stitt, M. (1998) Overexpression of pyrophosphatase leads to increased sucrose degradation and

starch synthesis, increased activities of enzymes for sucrose–starch interconversions, and increased levels of nucleotides in growing potato tubers. *Planta*, **205**, 428–437.

Godber, I.M. and Parsons, R. (1998) Translocation of amino acids from stem nodules of *Sesbania rostrata* demonstrated by GC–MS *in planta* ^{15}N isotope dilution. *Plant, Cell Environ.* **21**, 1089–1099.

Goddijn, O. and Smeekens, S. (1998) Sensing trehalose biosynthesis in plants. *Plant J.* **14**, 143–146.

Gonzalez, F.R. and Nardillo, A.M. (1999) Retention index in temperature-programmed gas chromatography. *J. Chromatogr. A.* **842**, 29–49.

Gzik, A. (1996) Accumulation of proline and pattern of alpha-amino acids in sugar beet plants in response to osmotic, water and salt stress. *Environ. Exp. Botany*, **36**, 29–38.

Hare, P.D., Cress, W.A. and van Staden, J. (1998) Dissecting the roles of osmolyte accumulation during stress. *Plant, Cell Environ.* **21**, 535–553.

Hare, P.D., Cress, W.A. and van Staden, J. (1999) Proline synthesis and degradation: a model system for elucidating stress-related signal transduction. *J. Exp. Botany*, **333**, 413–434.

Kamboj, J.S., Blanke, P.S., Quinlan, J.D. and Baker, D.A. (1999) Identification and quantification by GC–MS of zeatin and teantin riboside in xylem sap from rootstock and scion of grafted apple treas. *Plant Growth Regul.* **28**, 199–205.

Karakas, B., Ozias-Akins, P., Stushnoff, C., Suefferheld, M. and Rieger, M. (1997) Salinity and drought tolerance of mannitol-accumulating transgenic tobacco. *Plant, Cell Environ.* **20**, 609–616.

Katona, Z.F., Sass, P. and Molnar-Perl, I. (1999) Simultaneous determination of sugars, sugar alcohols, acids and amino acids in apricots by gas chromatography–mass spectrometry. *J. Chromatogr. A.* **847**, 91–102.

Merlo, L., Geigenberger, P., Hajirezaei, M. and Stitt, M. (1993) Changes of carbohydrates, metabolites and enzyme activities in potato tubers during development, and within a single tuber along a stolon–apex gradient. *J. Plant Physiol.* **142**, 392–402.

Morcuende, R., Krapp, A., Vaughan, H. and Still, M. (1998) Sucrose-feeding leads to an increased rates of nitrate assimilation, increased rates of α-oxoglutarate, and increased synthesis of a wide spectrum of amino acids in tobacco leaves. *Planta*, **206**, 394–409.

Müller, J., Wiemken, A. and Aeschbacher, R. (1999) Trehalose metabolism in sugar sensing and plant development. *Plant Sci.* **147**, 37–47.

Müller-Röber, B., Sonnewald, U. and Willmitzer, L. (1992) Inhibition of the ADP-glucose pyrophosphorylase in transgenic potatoes leads to sugar-storing tubers and influences tuber formation and expression of tuber storage protein genes. *EMBO J.* **11**, 1229–1238.

Murashige, T. and Skoog, F. (1962) A revised medium for rapid growth and bioassays with tobacco tissue cultures. *Physiol. Plant*, **15**, 473–497.

Prinsen, E., van Dongen, W., Esmans, E.L. and van Onckelen, H.A. (1998) Micro and capillary liquid chromatography–tandem mass spectrometry: a new dimension in phytohormone research. *J. Chromatogr. A.* **826**, 25–37.

Ramanjulu, S. and Sudhakar, C. (1997) Drought tolerance is partly related to amino acid accumulation and ammonia assimilation: a comparative study in two mulberry genotypes differing in drought sensitivity. *J. Plant Physiol.* **15**, 345–350.

Schweer, H. (1982) Gas chromatography–mass spectrometry of aldoses as *o*-methoxime, *o*-2-methyl-2-propoxime and *o*-*n*-butoxime pertriflouroacetyl derivatives on OV-225 with methylpropane as ionization agent. *J. Chromatogr.* **236**, 355–360.

Sonnewald, U., Hahirezaei, M.R., Kossmann, J., Heyer, A., Trethewey, R.N. and Willmitzer, L. (1997) Increased potato tuber size resulting from apoplastic expression of a yeast invertase. *Nature Biotech.* **15**, 794–798.

Sweetlove, L.J., Burrell, M.M. and ap Rees, T. (1996) Characterisation of transgenic potato (*Solanum tuberosum*) tubers with increased ADP glucose pyrophosphorylase. *Biochem. J.* **320**, 487–492.

Tanaka, Y., Kishimoto, Y., Otsuka, K. and Terabe, S. (1998) Strategy for selecting separation solutions in capillary electrophoresis–mass spectrometry. *J. Chromatogr. A.* **817**, 49–57.

Trethewey, R.N., Geigenberger, P., Riedel, K., Hajirezaei, M.R., Sonnewald, U., Stitt, . M., Riesmeier, J. and Willmitzer, L. (1998) Combined expression of glucokinase and invertase in potato tubers leads to a dramatic reduction in starch accumulation and a stimulation of glycolysis. *Plant J.* **15**, 109–118.

Trethewey, R., Geigenberger, P., Hennig, A., Fleischer-Notter, H., Müller-Röber, B. and Willmitzer, L. (1999) Induction of the activity of glycolytic enzymes correlates with enhanced hydrolysis of sucrose in the cytosol of transgenic potato tubers. *Plant, Cell Environ.* **22**, 71–79.

Veramendi, J., Willmitzer, L. and Trethewey, R. (1999a) *In vitro* grown potato microtubers are a suitable system for the study of transgenic lines altered in primary carbohydrate metabolism. *Plant Physiol. Biochem.* **37**, 1–5.

Veramendi, J., Roessner, U., Renz, A., Willmitzer, L. and Trethewey, R.N. (1999b) Antisense repression of hexokinase 1 leads to an accumulation of starch in leaves of transgenic potato plants but not to significant changes in tuber carbohydrate metabolism. *Plant Physiol.* **121**, 123–133.

Visser, R.G.F., Vreugdenhil, D., Hendriks, D. and Jacobsen, T. (1994) Gene expression and carbohydrate metabolism during stolon to tuber transition in potatoes (*Solanum tuberosum* L.). *Physiol. Plant.* **90**, 285–292.

Yamaguchi, I., Nakazawa, H., Nakagawa, R., Suzuki, Y., Kurogochi, S., Murofushi, N., . Takahashi, N. and Weiler, E.W. (1990) Identification and semi-quantification of gibberellins from the pollen and anthers of *Zea mays* by immunoassay and GC/MS. *Plant Cell Physiol.* **31**, 1063–1069.

PROTOCOL

Gas chromatography mass spectrometry–based metabolite profiling in plants

Jan Lisec[1,2], Nicolas Schauer[1,2], Joachim Kopka[1], Lothar Willmitzer[1] & Alisdair R Fernie[1]

[1]Max-Planck-Institut für Molekulare Pflanzenphysiologie, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany. [2]These authors contributed equally to this work. Correspondence should be addressed to A.R.F. (fernie@mpimp-golm.mpg.de).

Published online 27 June 2006; doi:10.1038/nprot.2006.59

The concept of metabolite profiling has been around for decades, but technical innovations are now enabling it to be carried out on a large scale with respect to the number of both metabolites measured and experiments carried out. Here we provide a detailed protocol for gas chromatography mass spectrometry (GC-MS)-based metabolite profiling that offers a good balance of sensitivity and reliability, being considerably more sensitive than NMR and more robust than liquid chromatography–linked mass spectrometry. We summarize all steps from collecting plant material and sample handling to derivatization procedures, instrumentation settings and evaluating the resultant chromatograms. We also define the contribution of GC-MS–based metabolite profiling to the fields of diagnostics, gene annotation and systems biology. Using the protocol described here facilitates routine determination of the relative levels of 300–500 analytes of polar and nonpolar extracts in ~400 experimental samples per week per machine.

INTRODUCTION

Although metabolite measurements have been carried out for decades owing to the fundamental regulatory importance of these molecules as components of metabolic pathways, the importance of some metabolites in the human diet and their use as diagnostic markers for a wide range of biological conditions, including disease and response to chemical treatment, is only now being recognized[1]. Historically, the measurement of metabolites was achieved either by spectrophotometric assays capable of detecting single metabolites or by simple chromatographic separation of mixtures of low complexity. Over the past decade, however, several methods offering both high accuracy and sensitivity for the analysis of highly complex mixtures of compounds have been established[2–8]. These methods include GC-MS, liquid chromatography mass spectrometry (LC-MS), capillary electrophoresis mass spectrometry (CE-MS) and Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS). In addition, chromatographically coupled NMR technologies have found great utility in addressing specific issues, particularly in the medical field[9,10] and perhaps most importantly with respect to the unequivocal determination of metabolite structures[11]. Nevertheless, NMR shows relatively low sensitivity and thus can be used for highly abundant metabolites when profiling complex mixtures.

GC-MS facilitates the identification and robust quantification of a few hundred metabolites in a single plant extract[7,8,12], resulting in fairly comprehensive coverage of the central pathways of primary metabolism. The main advantages of this technology are that it has long been used for metabolite profiling and thus there are therefore stable protocols for machine setup and maintenance, and chromatogram evaluation and interpretation. Although no single analytical system can cover the whole metabolome, GC-MS has a relatively broad coverage of compound classes[4], including organic and amino acids, sugars, sugar alcohols, phosphorylated intermediates and lipophilic compounds. Recovery experiments of all measurable classes of compounds have been done during method validation. For unknown compounds, recovery rates can be determined by recombination experiments in which extracts of two plant species are evaluated both independently and after mixing[13,14].

Although liquid chromatography–based methods offer distinct advantages, such as the broader range of metabolites detectable[2,15–17], they suffer from the lower reproducibility of retention times in liquid chromatography; in addition, owing to the predominant use of electron spray ionization, they are more susceptible to ion suppression effects, which render accurate quantification more difficult. Two alternative mass spectrometry technologies, FT-ICR-MS and CE-MS, are worth mentioning. FT-ICR-MS has unrivalled mass accuracy, thereby enabling the researcher to obtain directly a good idea about the chemical composition of the respective compound; however, a robust documentation of the validity of this technology, specifically with respect to quantification, is lacking for broad metabolite profiling. More data are available for CE-MS, a technology that detects low-abundance metabolites and affords good chromatographic separation. Despite robust validation of this procedure[6], however, only a few reports document its use[18,19].

Gas chromatography time-of-flight mass spectrometry (GC-TOF-MS) technology has been developed[20–22] and offers several advantages over the previously used quadrupole technology (GC-quad-MS)—notably fast scan times, which give rise to either improved deconvolution or reduced run times for complex mixtures and higher mass accuracy. For these reasons, the protocol described here is based on GC-TOF-MS technology; however, GC-quad-MS could be alternatively used in combination with published mass spectral alignment tools such as XCMS[23], MSFACTs[24], MetAlign (http://www.metalign.nl), AnalyzerPro (http://www.spectralworks.com) and BinBase (http://fiehnlab.ucdavis.edu). A detailed protocol for GC-quad-MS can be obtained by contacting the authors.

This protocol uses a MDN-35 or equivalent column with fatty acid methylesters (FAMEs) as retention time standards and was chosen because of the relative ease of application and fast chromatographic times[25,26]. *N*-Methyl-*N*-(trimethylsilyl)trifluoroacetamide (MSTFA) and methoxyamine hydrochloride are used as derivatization reagents because initial studies indicated that these compounds were the most appropriate for profiling of plant

metabolites[14,27]. Despite this, it has been shown that derivatization time and temperature influence the outcome of the results[28]. Derivatization of compounds often results in more than one peak for a metabolite of interest, owing to either partial silylation or isomerization in the case of methoxyaminated compounds such as sugars. In this protocol we identify all peaks of one compound, calculate their response independently, and pick the more reliable one using the statistical methods described here, but other methods such as summation of the peaks of one compound could be used as an alternative strategy. Further developments are ongoing and deal with this issue, in addition to degradation and partial silylation effects (J.K. and N.S., unpublished data).

Application of metabolite profiling

Improvements in metabolite profiling have rendered it an important tool for addressing biological problems. Previously, its main applications have been in the areas of diagnostics and descriptive analysis of metabolic response to various experimental perturbations, but increasingly examples of its use in gene function annotation and systems biology are being reported. Metabolite profiles have been widely used in conjunction with statistical tools for diagnosis: they have been used to infer the mode of action of various herbicides on barley seedlings[29], and to discriminate *Arabidopsis*, potato and tomato genotypes[7,8], various tissues of *Lotus japonicus*[30] and different stages of tomato fruit ripening[13].

In combination with a second round of experimental perturbation, diagnostic tools have been used to identify the principle metabolic change leading to metabolic shifts apparent after genetic perturbation[31,32]. Metabolite profiling has also been used in the process of testing whether genetically modified plants are substantially equivalent to conventional crops[33,34] and in understanding the complex shifts in metabolism that occur under nutrient limitation[35–37] and biotic stress[38,39]. Taken together, these examples show that metabolite profiling has important applications in the diagnostic characterization of different genetic and environmental conditions and can also aid in understanding the complex changes apparent under such circumstances.

In addition to its above-mentioned utility in diagnostics, metabolic profiling provides direct functional information on metabolic phenotypes and indirect information on a range of phenotypes that are determined by small molecules, such as stress tolerance or disease manifestations[1]. Given this, there is great potential for metabolite profiling as a tool for functional genomics. Indeed, gain-of-function analysis by the transgenomic expression of every gene of the *Escherichia coli* and yeast genomes independently in *Arabidopsis thaliana* both confirmed expected functions and facilitated the assignment of gene function to unannotated open reading frames[1]. This experiment was reliant on the fact that metabolite profiling can be used in a high-throughput format. Indeed, of all of the genomics technologies, it offers one of the best combinations of practical performance and cost per sample.

Metabolite profiling has also been used to demonstrate gene function by comparison of profiles derived from knockout mutants of *Arabidopsis* to their respective wild-type ecotypes, facilitating the annotation of genes associated with isoflavonoid, triterpenoid, pyridine alkaloid, glucosinolate, flavonoid and sterol metabolism[40–44]. The approach of focusing on individual genes can be extended to exploring the phenotypic relevance of genomic regions[45,46]. A GC-MS profiling study of breeding populations of tomato, wherein genomic sequences from the wild tomato species *Solanum pennellii* were introgressed into the elite cultivated species *Solanum lycopersicum*, identified nearly 900 quantitative trait loci for fruit metabolite accumulation and ultimately, through the study of progressively smaller recombinant introgressions, should facilitate the identification of genes that regulate metabolite content in a species of nutritional significance[46]. Similarly, the integration of metabolite and transcript profiling data has proved effective for identifying candidate genes for biotechnology[47,48].

In all technologies for metabolite profiling, the main limitation is the number of metabolites that can be detected and quantified. As ~200,000 metabolites are estimated to exist in the plant kingdom, it is clear that we are a long way from detecting the complement of plant small molecules. The availability of a full complement of isotopically labeled standards could greatly aid metabolite quantification, and further progress is undoubtedly required in determining the chemical identity of the peaks that can be resolved by current metabolite profiling methods. The use of metabolic profiling as a diagnostic tool is largely independent of the above-mentioned limitations, but its application to gene function analysis and systems biology depends largely on technological improvements. The fact that the phenotype of any biological system is largely dependent on its metabolite composition[1], however, gives ample reason to invest resources in attaining this goal. Although the protocol described here was developed for the analysis of *Arabidopsis* leaf samples, its use has been validated for plant heterotrophic tissues, highlighting its broad utility.

Considerations for the procedure

Metabolite profiling, like any technique concerned with measuring metabolites, requires the immediate inactivation of metabolism because the turnover of metabolites, as compared with proteins and DNA or RNA, is extremely rapid. Quenching of metabolism is generally achieved by rapidly freezing samples (at a constant temperature of −60 °C or less). In addition, the whole procedure critically requires materials of the highest purity to prevent contamination, which can easily influence the outcome of the experiment. Given this hazard, it is necessary to run quality control samples frequently alongside each experiment. Basic requirements for an experiment should be considered *a priori* for a generalized standard design[49]. The following details are critically important. Given that experiments generally comprise sets of samples of interest and their respective controls, it is a prerequisite that these samples are comparable to one another. For large sample sets it is imperative that there are a sufficient number of control samples, particularly because it is sometimes not possible to measure all samples in a single GC-MS run. Alongside each experiment, blank samples should be run for to identify contaminants. Blank samples should be derivatized alongside the other samples—the only difference is that this sample vial contains no metabolite extract. Another important detail is the reproduction of biological data—a minimum number of six biological replicates is sufficient[7], but where possible the number of replicates should be even higher, especially if in-depth statistical analysis of the data sets emerging from the analysis is intended[50–52]. Indeed, consideration of elementary statistic suggests that the sample number requirements can be determined by power analysis determined from the degree of variance within populations.

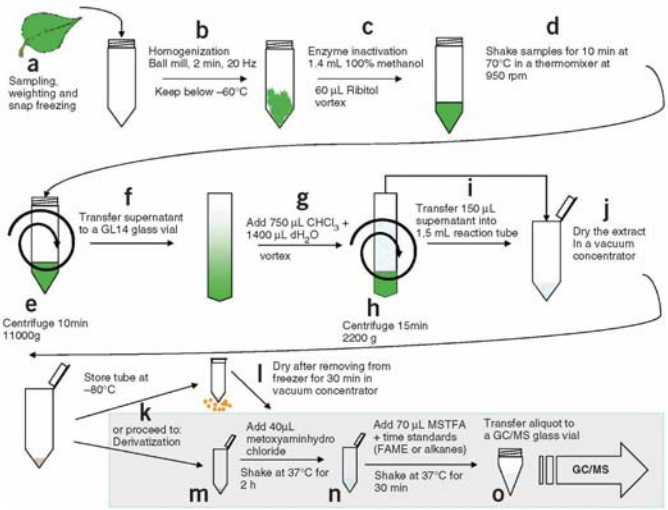


Figure 1 | Experimental procedure for extract preparation. (**a**) Sampling, weighting and snap-freezing. (**b**) Homogenization in a ball mill for 2 min at 20 Hz. Keep the temperature below −60 °C. (**c**) Enzyme inactivation. Add 1.4 ml of 100% methanol, vortex sample, add 60 µl of Ribitol, and vortex sample. (**d**) Shake samples for 10 min at 70 °C in a thermomixer at 950 r.p.m. (**e**) Centrifuge sample for 10 min at 11,000*g*. (**f**) Transfer supernatant to a glass vial. (**g**) Add 750 µl of chloroform and 1,400 µl of dH_2O to the sample, and vortex. (**h**) Centrifuge sample for 15 min at 2,200*g*. (**i**) Transfer 150 µl of supernatant into a 1.5-ml reaction tube. (**j**) Dry the extract in a vacuum container. (**k**) Store the tube at −80 °C or proceed to derivatization. (**l**) If storing the tube, dry the sample in a vacuum concentrator for 30 min after removing from the freezer. (**m**) Add 40 µl of methoxyamination reagent (see REAGENT SETUP) to sample and shake at 37 °C for 2 h. (**n**) Add 70 µl MSTFA reagent and time standards (such as FAMEs or alkanes) to sample, and shake at 37 °C for 30 min. (**o**) Transfer aliquot to a GC-MS glass vial and analyze by GC-MS.

Sampling and extraction

Before beginning the sampling process, the time point for sampling must be carefully considered. As a general rule, we harvest photosynthetic leaf tissue in the middle of the light period because experiments in our own laboratory have indicated that almost all metabolites that we can detect and quantify are subject to strong diurnal rhythms[53]. We also tend to take samples from plants before emergence of the first inflorescence, always harvesting from the same internode and using fully developed, nonsenescent leaves. Experience dictates that these factors are crucial[30,54,55], as is the rapidity of the process because many metabolites show turnover times of a fraction of a second[56]. Although those metabolites that can be readily detected by GC-MS methods generally turnover less quickly, rapid quenching is still critical[2].

Cut samples should be rapidly weighed and then immediately frozen to quench metabolism. Before homogenization of the sample, all laboratory material to be used should be cooled down to prevent thawing of the biological sample. The amount of tissue (100 mg) used in this protocol differs from those used by other groups, but the solvent-to-tissue ratio is conserved[28]. The main reason for taking samples of relatively high mass is that lesser amounts are more difficult to handle and small errors in weighing can propagate to produce large changes in the final evaluation. If the user chooses to use less tissue, however, the extraction volume can be readily adapted. The first three steps of the extraction procedure (**Fig. 1**) are particularly crucial in terms of avoiding thawing and its associated problems.

This protocol is suitable for both polar and apolar extraction of metabolites. Although there is considerable experience in the analysis of polar metabolites, far less is known about apolar compounds, owing, at least in part, to carryover and contamination effects, which require more sophisticated knowledge, equipment and methodology. For this reason, here we concentrate only on the polar phase. Although we supply precise information pertaining to the instrumentation used, it should be noted that this protocol is broadly applicable to all machines of this type.

MATERIALS

REAGENTS

- Argon 5.0 (Messer-Griesheim)
- Chloroform for liquid chromatography (Merck, cat. no. 67-66-3) **! CAUTION** Chloroform is toxic and should be handled under a fume hood
- Helium 5.0 carrier gas (Air Liquide)
- Methanol, gradient grade for liquid chromatography (Merck, cat. no. 67-56-1)
- *N*-Methyl-*N*-(trimethylsilyl)trifluoroacetamide (MSTFA reagent; Macherey-Nagel, cat. no. 24589-78-4); prepared in 1-ml vials and stored at 4 °C. **! CAUTION** Reagent is extremely toxic and should be handled under the fume hood
- Orange silica gel, no. 77.1 (Carl Roth)
- Ribitol (Sigma, cat. no. 488-81-3)

EQUIPMENT

- Autosampler and software (CTC Combi PAL and PAL cycle composer software version 1.5.0; CTC Analytics); the configuration comprises an agitator-incubator oven, a 98-sample tray for 2.0-ml vials, a 32-sample tray for 10–20-ml vials, three 100-ml solvent reservoirs (i.e. a syringe wash station and a liquid version 25-µl syringe kit mounted on the robotic autosampler arm)
- Conical single taper split/splitless liner (Agilent)
- Gas chromatograph, 6890N, split/splitless injector with electronic pressure control up to 150 psi (Agilent)
- GL14 glass vials (Schott)
- MDN-35 capillary column, 30-m length, 0.32-mm inner diameter, 0.25-µm film thickness (e.g. Macherey-Nagel or equivalent[57])
- Micro-vials: 1.5 ml, safe-lock, tapered bottom, and 2.0 ml, screw-cap, round bottom (Eppendorf)
- Oscillating ball mill, MM200 (Retsch)
- Pegasus III time-of-flight mass spectrometer (Leco Instruments)
- Steel balls, VA5mm (Th. Geyer Berlin)
- Screw caps for GL14 glass vials (Schott, cat. no. 29 990 12 04)
- Teflon adaptor for 1.5–2.0 ml micro-vials (Retsch)
- Automated mass spectral deconvolution and identification system (AMDIS; National Institute of Standards and Technology)

PROTOCOL

- ChromaTOF chromatography processing and mass spectral deconvolution software, version 1.00 or higher, driver 1.61 or higher (LECO Instrumente), running on a state-of-the-art computer with a minimum of 512-MB RAM and an 1.0G-Hz Pentium IV processor or equivalent
- R: a Language and Environment for Statistical Computing (R Foundation for Statistical Computing)

REAGENT SETUP

Methoxyamination reagent Dissolve methoxyamine hydrochloride (Sigma, cat. no. 593-56-6) at 20 mg ml^{-1} in pure pyridine (Merck, cat. no. 110-86-1) at 20–25°C. This reagent needs to be prepared freshly before the experiment. ! CAUTION Reagents are extremely toxic and should be handled under the fume hood.

Retention time index standard mixture Dissolve FAMEs in chloroform at a final concentration of 0.4 ml ml^{-1} or 0.8 mg ml^{-1} for liquid or solid standards. Reagent can be stored at −4 °C. Esters included are methylcaprylate (Sigma, cat. no. 111-11-5), methyl pelargonate (Sigma, cat. no. 1731-84-6), methylcaprate (Sigma, cat. no. 110-42-9), methyllaurate (Sigma, cat. no. 111-82-0), methylmyristate (Sigma, cat. no. 124-10-7), methylpalmitate (Sigma, cat. no. 112-39-0), methylstearate (Sigma, cat. no. 112-61-8), methyleicosanoate (Sigma, cat. no. 1120-28-1), methyldocosanoate (Sigma, cat. no. 929-77-1), lignoceric acid methylester (Sigma, cat. no. 2442-49-1), methylhexacosanoate (Sigma, cat. no. 5802-82-4), methyloctacosanoate (Sigma, cat. no. 55682-92-3), and triacontanoic acid methylester[25] (Sigma, cat. no. 629-83-4). Alternatively, alkanes[13] or fatty acids[7] have been, and can be, used.

EQUIPMENT SETUP

Standardization As a rule, metabolite profiling studies compare two or more states of a given biological system; thus, absolute quantification is not necessary and relative quantifications of the level of metabolites of interest per tissue mass (i.e. per gram of fresh weight) is sufficient. In such instances, the challenge of quantification is reduced to comparison between one or many samples, which essentially transforms the problem of quantification into a problem of standardization. Any standardization has to correct for the following. (1) Experimental errors during sample preparation (determination of the sample amount and subsequent liquid handling): this is corrected for by the critical inclusion of a compound not present in biological samples (i.e. Ribitol or ^{13}C Sorbitol), directly after homogenization of the sample. For normalization with Ribitol, the unique mass *m/z* 319 is used, whereas for ^{13}C Sorbitol *m/z* 323 is used. (2) Overall machine sensitivity. (3) Changes in sensitivity towards specific compounds owing to differences in the matrix. Although this aspect in the form of ion suppression is a major problem in any mass spectrometry technique relying on electrospray ionization, as a rule it is a lesser problem in GC-coupled mass spectrometers.

Machine sensitivity Given the variability in the overall machine, sensitivity is most probably the crucial factor to correct for during quantification procedures. In principle, the following standardization protocols can be applied to correct for machine sensitivity. (1) The expression of every identifier mass trace used for quantification as a proportion of the total ion intensity of all identified compounds of that sample. In our experience this is only a very gross correction for machine performance and not overly reliable. (2) The evaluation of a large number of controls (~20% of the total samples) in a random order between experimental samples is highly preferable. This control should be as similar as possible with respect to chemical complexity to the experimental samples. In practice, this means that one need only to prepare a large batch of pooled extracts from, for example, leaves of a given species and aliquot these for subsequent use as machine sensitivity controls. This quantification is the most reliable; however, it requires a large number of control samples to be run and thus leads to increased costs and reduced throughput. An alternative that is used in experiments where absolute quantification is a prerequisite, such as analyses of metabolic fluxes[58,59], is the evaluation of calibration curves of authentic standards. Such standards are analyzed in replicate over a dilution series after derivatization and handling in the exact same way as described above. This approach is also cost- and labor-intensive but offers the advantage of facilitating comparison with published metabolite data obtained with other analytic techniques[60]. (3) Given that machine sensitivity is generally sufficiently stable over a day, the median of distribution of each metabolite across the samples measured in a day can be calculated and the content of each metabolite can be subsequently expressed in comparison to its daily median. This is a very cost-effective approach because it does not require as many controls to be run as in the previous approach. In addition, it allows a metabolite-by-metabolite correction for machine sensitivity. An example of this approach is given in **Figure 2**: measurement of glycine over five independent days shows some variance; when samples are related to the daily median, however, the variance between biological replicates is comparable between measurement days. It is important to note that this approach is valid only when the chemical composition is very similar and the distributions of concentration in the different samples are similar in the samples measured on different days. If these prerequisites are fulfilled, this approach is both a robust and reliable one.

PROCEDURE

Sampling and extraction

1| Sample leaf material in 2-ml, screw cap, round bottom tubes. Define the exact mass of plant sample (~100 mg of fresh weight) and rapidly freeze the sample-containing vial using liquid nitrogen or an equivalent low-temperature liquid.

2| To homogenize the tissue, place steels balls into the sample tubes and insert samples into precooled Teflon adaptors. Homogenize in ball mill for 2 min at 20 Hz.

■ PAUSE POINT Frozen homogenate can be stored at −80 °C for up to 3 months.

3| Add 1,400 µl of 100% methanol (pre-cooled at −20 °C) and vortex for 10 s.

4| Add 60 µl of Ribitol (0.2 mg ml^{-1} stock in dH_2O) as an internal quantitative standard and vortex for 10 s.

5| Shake for 10 min at 70 °C in a thermomixer at 950 r.p.m.

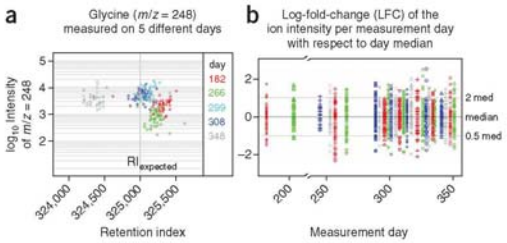


Figure 2 | Normalization. Although we convert the data from each sample individually from retention time to RI, there is still daily variation over long measurement periods. Detector sensitivity is another factor that we have to take into account to enable us to perform large-scale experiments. (**a**) Glycine (m/z = 248) measured on five different days: intensity is clearly dependent on the day of measurement (samples came from the same experiment and were completely randomized). (**b**) Log fold-change of the ion intensity per measurement day with respect to day median: if we normalize our results on the median value for a specific metabolite per measurement day, the distribution of the samples is comparable.

6| Centrifuge for 10 min at 11,000*g*.

7| Transfer supernatant to a Schott GL14 glass vial.

8| Add 750 µl of chloroform (−20 °C).

9| Add 1,500 µl dH_2O (4 °C) and vortex for 10 s.

10| Centrifuge 15 min at 2,200*g*.

11| Transfer 150 µl from the upper phase (polar phase) into a fresh 1.5-ml tube.

12| As a backup (in case you lose a sample), take a second aliquot into a new 1.5-ml tube.

13| Dry in a vacuum concentrator without heating.

14| Before freezing the aliquots at −80 °C, fill the tubes with argon gas and place them inside a plastic bag containing silica bead desiccant. Argon-filled sample vials prevent the extract from oxidization and degradation by reactions through components of atmospheric air.
! CAUTION Halogenic reagents and solutions should be disposed with halogenic waste.
■ PAUSE POINT Samples can be stored at −80 °C for up to 3 months.

Derivatization

15| Place samples stored at −80 °C in a vacuum concentrator for 30 min before derivatization.

16| Add 40 µl of methoxyamination reagent (see REAGENT SETUP) to the aliquots.
! CAUTION Derivatization reagents are extremely toxic. Handle with absolute care. Work with gloves and under the fume hood.
▲ CRITICAL STEP In the process of derivatization, condensation of reagents appears on the wall and lid of the reaction tubes; therefore, centrifugation of the reaction mixture is essential after every incubation step.

17| Also prepare one derivatization reaction using an empty reaction tube as a control.

18| Shake for 2 h at 37 °C.

19| Prepare MSTFA reagent with 20 µl ml^{-1} of retention time index standard mixture (see REAGENT SETUP).

20| Add 70 µl of the solution prepared in Step 19 to the sample aliquots.

21| Shake for 30 min at 37 °C.

22| Transfer into glass vials suitable for GC-MS analysis.
▲ CRITICAL Steps 15–22 have been shown to be very critical. In this protocol we use derivatization reagent in supersaturated concentrations to ensure completion of derivatization.

GC-TOF/MS metabolite profiling: Injection parameters

23| Inject 1 µl of sample at 230 °C in splitless mode with helium carrier gas flow set to 2 ml min^{-1} by using the autosampler setup (see EQUIPMENT) The flow rate is kept constant with electronic pressure control enabled. Optionally, but especially recommended in cases of high metabolite concentrations, injection can be done in split mode with the split ratio adjusted to 1:25.

GC-TOF/MS metabolite profiling: Chromatography parameters

24| Perform chromatography with a 30-m MDN-35 capillary column. The temperature program should be isothermal for 2 min at 80 °C, followed by a 15 °C per min ramp to 330 °C, and holding at this temperature for 6 min. Cooling should be as rapid as instrument specifications allow. Set the transfer line temperature to 250 °C and match ion source conditions[57].

GC-TOF/MS metabolite profiling: Mass Spectrometer parameters

25| Set the ion source to maximum instrument specifications, 250 °C. The recorded mass range should be *m/z* 70 to *m/z* 600 at 20 scans per s. Proceed the remaining monitored chromatography time with a 170-s solvent delay with filaments turned off. Manual mass defect should be set to 0, filament bias current should be −70 V, and detector voltage should be ~1700–1850 V. Automatically tune the instrument according to the manufacturer's instructions.

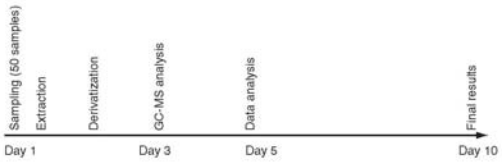


Figure 3 | Timeline of standard operating procedure. Note that five days are assigned to final results despite the use of an instantaneous algorithm, because manual inspection of chromatograms is a highly advisable quality control.

ANTICIPATED RESULTS

Figure 3 visualizes the estimated amount of time from collecting 50 biological samples to final results. Sampling, extraction and derivatization takes less than 50% of the time, but more effort and time, approximately between 2 and 5 days, is needed for the comprehensive data analysis.

Deconvolution

Following the general philosophy of metabolic profiling, the extraction procedure used should introduce as little bias as possible with respect to the complexity of the compounds extracted from the biological sample. Thus, metabolic profiling leads to complex chromatograms characterized by coeluting compounds and vast differences in the relative abundance of the different compounds. Although problematic, these issues can be partially resolved by deconvolution of the chromatograms. The machine supplier's software, e.g. ChromaTOF, offers a build-in deconvolution algorithm. Deconvoluted spectra can be exported as plain text files for further processing. We suggest the following parameters for the deconvolution process (with acceptable range in parentheses). In all instances we used the machine manufacturer's recommended approaches, which we have found to be highly appropriate: Baseline offset = 1 (0.5–1); Smoothing = 5 data points (3–7); Peak width = 3 s (3–4 s); *S/N* (signal-to-noise ratio) = 10 (2–15).

Retention time index

Retention time index (RI) is probably the most important parameter for peak assignment. In our experience it is absolutely crucial that each chromatogram is corrected for retention times separately, as even within a day absolute retention times show variance that, combined with the fact that the complex mixtures apparent in plant extracts result in highly complex chromatograms, can lead to false peak annotations (see **Fig. 4** for an example of the variation of retention times of the FAME retention time standards). To minimize this problem, we apply an algorithm (R-Script 1; available from J.L.) comprising the following steps.

(i) Identification of the retention time for each of the internal markers (see 'REAGENT SETUP for internal retention time standards) and assign a 'fixed RI' to the respective peaks.
(ii) Calculation of the RI for all compounds eluting between two standards using a linear interpolation.
(iii) Extension of the linear correction for all compounds eluting prior to the first or after the last standard.

To be able to narrow considerably the time window to search for a certain compound, each file needs to be corrected independently. Although this can be theoretically achieved by the original software, it is highly impractical on a high-throughput scale.

Peak annotation

The R-script 1 algorithm developed in our laboratories facilitates annotation of a given peak to a compound (with known or unknown chemical structure), a process that is reliant on two known factors, namely the RI and the mass spectrum.

Unique masses as identifiers

In principle, it should be possible to annotate each compound based on its unique mass spectrum and RI (refs. 21, 61, 62). In metabolite profiling, however, the presence of coeluting compounds present in high dynamic range[4] can mean that reliance on these parameters proves to be difficult. This is even more pronounced when the coeluting compounds have one or more masses in common. Many commercial and publicly available mass spectral evaluation tools exist. These tools are largely similar in function, if not execution, and offer distinct

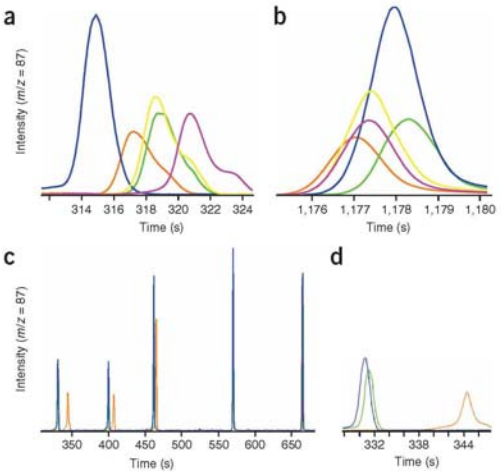


Figure 4 | Retention time variation within sample sets. (**a**,**b**) General variation of retention time of RI standards within a day. Generally we observe stronger variation (up to 6 s) for early eluting standards (**a**), whereas later standards are robust (**b**). (**c**,**d**) Outlier behavior at early elution times. Outliers may occur in the early elution phase, showing differences of up to 12 s if compared with other files from the same day, but lining up with those during the chromatographic run. Intensities of specific FAMEs of the retention index standards within a batch of chromatograms are shown. Selected files of an authentic data set are used for illustrative purposes. Different colors represent different exemplary data sets.

advantages and disadvantages. Because a broad comparison of these various algorithms is currently lacking, we do not discuss them in detail here, but rather concentrate on an algorithm developed specifically for the metabolite profiling method that we describe. For this purpose, we decided to use a combination of a very precise relative retention time as described above and one or more mass traces unique within this retention time window for the assignment of a given peak to a compound (R-Script 1). The RI-corrected spectra are processed by a second bespoke R-Script (R-Script 2, available from J.L.) according to a prepared reference list using an algorithm that achieves the following: all peaks within the specified time window are evaluated; and the peak showing the maximum intensity for the predefined unique ion is chosen.

The reference list—containing name, expected RI, allowed RI variation and unique mass for a number of metabolites—can be initially prepared by evaluating the GC-MS spectra resulting from the evaluation of a mixed sample pooled from aliquots of the whole measurement set in conjunction with available and constantly expanding GC-MS library sets[21,61,62], when necessary following the troubleshooting procedure to ensure authenticity of peak identification.

Identification of novel compounds and contaminants

Compound identification is essentially performed by running authentic standards and determining RI and specific masses. In many cases, however, specifically in the case of unknown

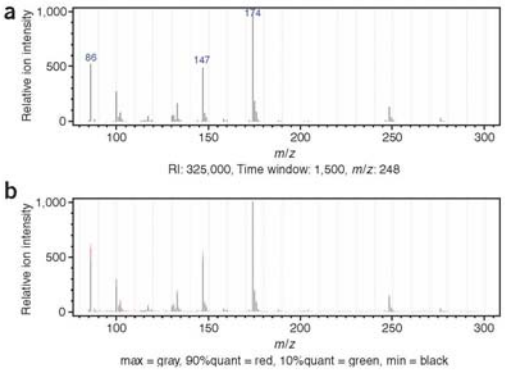
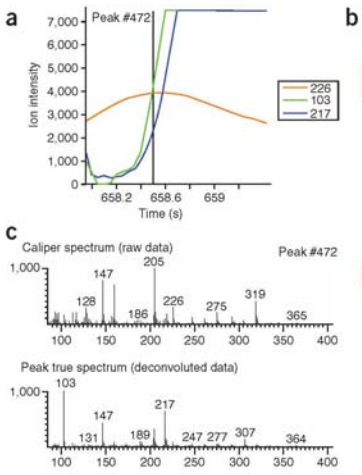


Figure 6 | Median and box-plot spectra. These plots are regularly used to check the quality of search results quickly. Here glycine was searched for with the following reference conditions and data set: RI, 325,000; time window, ±1,500 RI units (~1.5 s); unique mass, 248; data set, measurement day 182, 43 samples, no missing values. (**a**) Median value for each ion calculated from all 43 relative spectra chosen. It should be close to the recorded library spectrum for this metabolite. (**b**) Box-plot of the candidate spectra, indicating by color minimum, 10% quantile, median, 90% quantile and maximum of all extracted values. In this example, one could observe the biggest (but negligible) variation for the ion of mass 86. Only if this box-plot spectrum reveals strong variation between the chosen spectra do we calculate the mean squared difference for all single spectra from the median spectrum, thus quickly identifying the outliers, plotting them as an overview plot (see **Fig. 5**) and then re-evaluating the chromatogram where necessary.



Mass	Caliper	Peak true
103	2028	49843
217	1229	31597
147	10906	21370
205	14034	15716
204	2079	7222
218	60	6905
307	0	6567
219	2359	5378
189	659	4671
206	2825	4578
104	410	4446
148	2214	4356
160	9740	4351
226	4040	3889
191	182	3808
84	147	3107
113	2637	2972
277	459	2633
275	2761	2394
207	1268	2304
93	2870	2083
173	38	2063
149	666	2030
247	1566	2001
308	0	1896
131	1207	1870

Figure 5 | Example of a peak deconvolution error. (**a**) Unique mass and two high-abundance ion intensities from a coeluting peak. (**b**) Raw data (Caliper, the triangular visible at the base of **a**, indicates a single data point on the timescale, in this instance, identical to the time point of the deconvoluted peak) and deconvoluted data (Peak true) at peak position; ions from **a** are boxed. (**c**) Spectrum representation of raw and deconvoluted data at peak position. Note that deconvoluted data obviously do not always represent actual values. Spectral comparison would fail to identify this peak because it will be highly variable owing to deconvolution errors. Nevertheless, intensity values for the unique mass are sensible.

metabolites for which some mass spectral properties are clear, it is highly desirable to obtain a mass spectrum as an aid for their further identification. Median spectra (as used for error correction) may be computed from a number of samples and can be exported in NIST format for comparison to external databases. Given that many such databases report data for nonderivatized metabolites, the derivatization and subsequent analysis of authentic standards represents one way to identify unknown peaks via GC-MS. Other ways to tackle this difficult problem are largely reliant on analytical techniques such as LC-MS[63] and NMR; the exceedingly high mass accuracy of this technology, when coupled to chromatography, will have a considerable role in the future development of plant metabolomics.

The direct experimental output from this protocol is a list of metabolite contents of the experimental conditions in comparison to the control. The number of compounds detected in polar leaf extracts depends on the RI and mass spectral information annotated by the experimenter. In general, taking mass spectral tag information from publicly available libraries[21,61,62] into account, this should lead to ~150–500 compounds of known and unknown origin; however this number varies on the species and tissue type. The direct output described above can be subsequently evaluated with respect to the biological question of the research either in a metabolite-by-metabolite manner or by using one of the many available

statistical packages for multivariate analysis. The former analysis would be suggested in studies of metabolic regulation, for example coordinated metabolic responses to nutrient deprivation and for gene annotation, whereas the later retains great utility in diagnostics-based approaches. As a general rule, only 40% of the compounds are annotated to a specific metabolite, so if a particularly interesting trend in an unannotated metabolite is found in an experiment, it is recommended that a second analytical technique is used to determine the chemical structure of this unknown metabolite. A note of caution is necessary here, however, because this is generally a far from trivial task.

Deconvolution problems

Using the manufacturer's deconvolution software we have, in a few instances, encountered errors that can be defined either as errors of multiple deconvolution of a single peak, or as errors in which deconvoluted spectra contain the wrong ion intensities (**Fig. 5**). Although these errors occur infrequently, their frequency is high enough to preclude over-reliance on the machine manufacturer's own software. For this reason, we recommend that correction of these errors be carried out by collection all data files of a sample set (generally 40–50 samples) and processing these files together in the framework of the open source software package R using the designed scripts (available from J.L.)

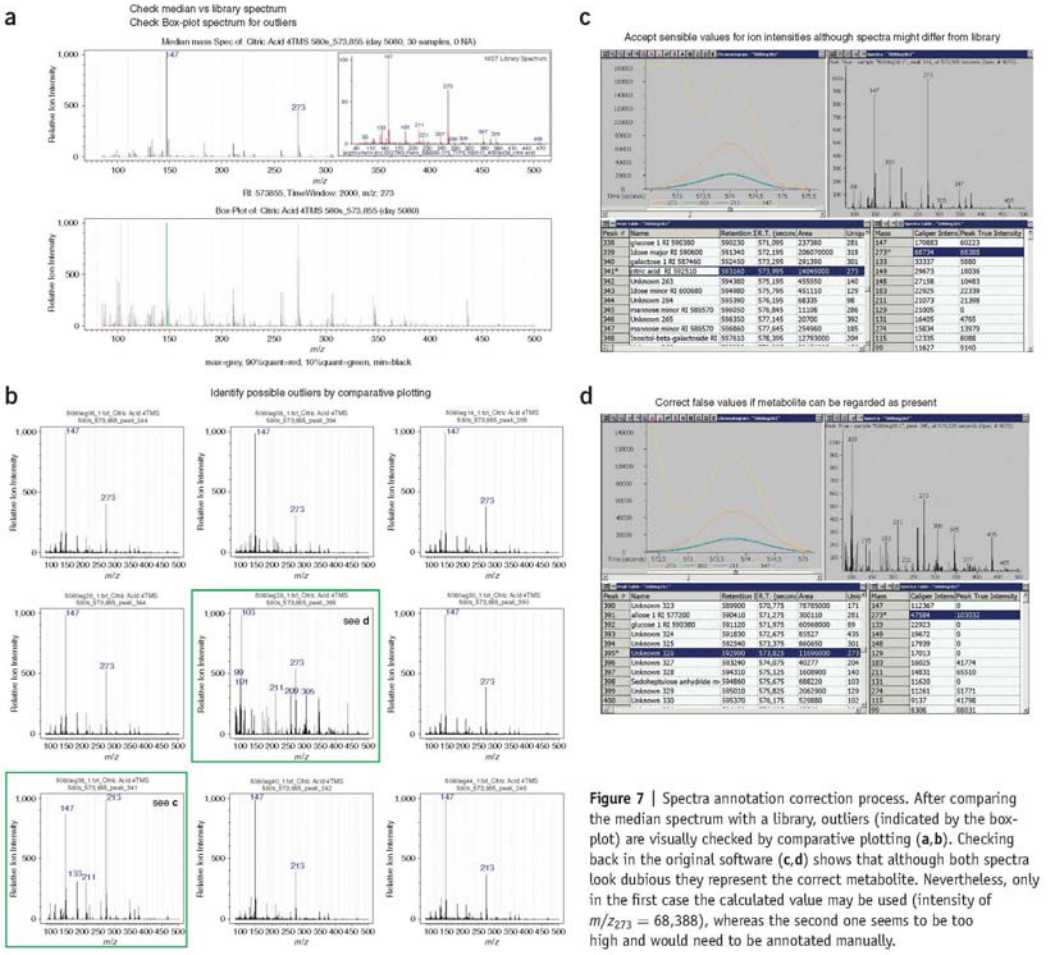


Figure 7 | Spectra annotation correction process. After comparing the median spectrum with a library, outliers (indicated by the box-plot) are visually checked by comparative plotting (**a**,**b**). Checking back in the original software (**c**,**d**) shows that although both spectra look dubious they represent the correct metabolite. Nevertheless, only in the first case the calculated value may be used (intensity of $m/z_{273} = 68{,}388$), whereas the second one seems to be too high and would need to be annotated manually.

Annotation

Within a given matrix (e.g. *Arabidopsis* leaf or root, potato tuber, tomato pericarp tissue, *Lotus japonicus* nodule), the above procedure for compound annotation is highly reliable. When a new matrix is analyzed or when a given matrix is analyzed where the organism was exposed to widely different environmental condition or a genetic variant shows a marked visual phenotype, we strongly suggest a manual inspection of chromatograms to avoid the erroneous use of a unique mass for compound identification owing to the appearance of a novel compound with the same mass trace in the same retention time window. Manual inspection is a highly time-consuming and laborious process. To speed up the process of error identification, however, a graphical overview of all analyzed spectra for a specified metabolite per data set can be used, as exemplified in **Figure 6**. Such a display can indicate dubiously annotated spectra. The box-plot spectra (**Fig. 6b**), represents the statistical comparison of the samples of a given set of chromatograms to the median standard spectra (**Fig. 6a**). Where variance is great across the chromatograms, additional plotting possibilities enable the user to trace back to the offending samples or peaks, thus facilitating a rapid verification or falsification of peak annotation and an acceleration of the process of manual correction. Only when the data are validated either algorithmically or manually are they deemed acceptable for publication and or storage in publicly accessible databases. An overview of the complete process is outlined in **Figure 7**.

ACKNOWLEDGMENTS We thank O. Fiehn and Ä. Eckhardt for intensive collegial discussions on all matters metabolomic.

COMPETING INTERESTS STATEMENTS The authors declare that they have no competing financial interests.

Published online at http://www.natureprotocols.com
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Fernie, A.R., Trethewey, R.N., Krotzky, A.J. & Willmitzer, L. Innovation—metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.* **5**, 763–769 (2004).
2. Kopka, J., Fernie, A., Weckwerth, W., Gibon, Y. & Stitt, M. Metabolite profiling in plant biology: platforms and destinations. *Genome Biol.* **5**, 109 (2004).
3. Hirai, M.Y. & Saito, K. Post-genomics approaches for the elucidation of plant adaptive mechanisms to sulphur deficiency. *J. Exp. Bot.* **55**, 1871–1879 (2004).
4. Sumner, L.W., Mendes, P. & Dixon, R.A. Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* **62**, 817–836 (2003).
5. Harrigan, G.G. & Goodacre, R. *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis* (Springer, Berlin/Heidelberg, 2003).
6. Soga, T. *et al.* Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *J. Proteome Res.* **2**, 488–494 (2003).
7. Roessner, U. *et al.* Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* **13**, 11–29 (2001).
8. Fiehn, O. *et al.* Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* **18**, 1157–1161 (2000).
9. Wasim, M., Hassan, M.S. & Brereton, R.G. Evaluation of chemometric methods for determining the number and position of components in high-performance liquid chromatography detected by diode array detector and on-flow H-1 nuclear magnetic resonance spectroscopy. *Analyst* **128**, 1082–1090 (2003).
10. Lindon, J.C. HPLC-NMR-MS: past, present and future. *Drug Discov. Today* **8**, 1021–1022 (2003).
11. Meiler, J. & Will, M. Genius: a genetic algorithm for automated structure elucidation from C-13 NMR spectra. *J. Am. Chem. Soc.* **124**, 1868–1870 (2002).
12. Halket, J.M. & Zaikin, V.G. Derivatization in mass spectrometry. 1. Silylation. *Eur. J. Mass Spectrom.* **9**, 1–21 (2003).
13. Roessner-Tunali, U. *et al.* Metabolic profiling of transgenic tomato plants overexpressing hexokinase reveals that the influence of hexose phosphorylation diminishes during fruit development. *Plant Physiol.* **133**, 84–99 (2003).
14. Roessner, U., Wagner, C., Kopka, J., Trethewey, R.N. & Willmitzer, L. Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J.* **23**, 131–142 (2000).
15. Swart, P.J. *et al.* HPLC-UV atmospheric-pressure ionization mass-spectrometric determination of the dopamine-D2 agonist N-0923 and its major metabolites after oxidative-metabolism by rat-liver, monkey liver, and human liver-microsomes. *Toxicol. Methods* **3**, 279–290 (1993).
16. Aharoni, A. *et al.* Nontargeted metabolome analysis by use of Fourier transform ion cyclotron mass spectrometry. *OMICS* **6**, 217–234 (2003).
17. Plumb, R.S. *et al.* Use of liquid chromatography/time-of-flight mass spectrometry and multivariate statistical analysis shows promise for the detection of drug metabolites in biological fluids. *Rapid Commun. Mass Spectrom.* **17**, 2632–2638 (2003).
18. Sato, S., Soga, T., Nishioka, T. & Tomita, M. Simultaneous determination of the main metabolites in rice leaves using capillary electrophoresis mass spectrometry and capillary electrophoresis diode array detection. *Plant J.* **40**, 151–163 (2004).
19. Unger, M. *et al.* Analytical characterisation of crude extracts from an African *Ancistrocladus* species using high-performance liquid chromatography and capillary electrophoresis coupled to ion trap mass spectrometry. *Phytochem. Anal.* **15**, 21–26 (2004).
20. Taylor, J., King, R.D., Altmann, T. & Fiehn, O. Application of metabolomics to plant genotype discrimination using statistics and machine learning. *Bioinformatics* **18**, S241–S248 (2002).
21. Wagner, C., Sefkow, M. & Kopka, J. Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry* **62**, 887–900 (2003).
22. Saito, K., Dixon, R.A. & Willmitzer, L. *Plant Metabolomics* (eds Nagata, T.L.H. & Widholm, J.M.) (Springer, Berlin/Heidelberg, 2006).
23. Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **78**, 779–787 (2006).
24. Duran, A.L., Yang, J., Wang, L.J. & Sumner, L.W. Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics* **19**, 2283–2293 (2003).
25. Weckwerth, W., Loureiro, M.E., Wenzel, K. & Fiehn, O. Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl. Acad. Sci. USA* **101**, 7809–7814 (2004).
26. Kaplan, F. *et al.* Exploring the temperature-stress metabolome of *Arabidopsis*. *Plant Physiol.* **136**, 4159–4168 (2004).
27. Fiehn, O., Kopka, J., Trethewey, R.N. & Willmitzer, L. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal. Chem.* **72**, 3573–3580 (2000).
28. Gullberg, J., Jonsson, P., Nordstrom, A., Sjostrom, M. & Moritz, T. Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of *Arabidopsis* thaliana samples in metabolomic studies with gas chromatography/mass spectrometry. *Anal. Biochem.* **331**, 283–295 (2004).
29. Sauter, H., Lauer, M. & Fritsch, H. Metabolic profiling of plants—a new diagnostic technique. *Am. Chem. Soc. Symp. Ser.* **443**, 288–299 (1991).
30. Desbrosses, G.G., Kopka, J. & Udvardi, M.K. *Lotus japonicus* metabolic profiling. Development of gas chromatography–mass spectrometry resources for the study of plant-microbe interactions. *Plant Physiol.* **137**, 1302–1318 (2005).
31. Roessner, U., Willmitzer, L. & Fernie, A.R. High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol.* **127**, 749–764 (2001).
32. Junker, B.H. *et al.* Temporally regulated expression of a yeast invertase in potato tubers allows dissection of the complex metabolic phenotype obtained following its constitutive expression. *Plant Mol. Biol.* **56**, 91–110 (2004).
33. Catchpole, G.S. *et al.* Hierarchical metabolomics demonstrates substantial compositional similarity between genetically modified and conventional potato crops. *Proc. Natl. Acad. Sci. USA* **102**, 14458–14462 (2005).
34. Defernez, M. *et al.* NMR and HPLC-UV profiling of potatoes with genetic modifications to metabolic pathways. *J. Agric. Food Chem.* **52**, 6075–6085 (2004).

35. Hirai, M. *et al.* Transcriptome and metabolome analyses reveal a whole adaptive process of plant to sulfur deficiency. *Plant Cell Physiol.* **45**, S122–S122 (2004).
36. Nikiforova, V.J. *et al.* Systems rebalancing of metabolism in response to sulfur deprivation, as revealed by metabolome analysis of *Arabidopsis* plants. *Plant Physiol.* **138**, 304–318 (2005).
37. Urbanczyk-Wochniak, E. & Fernie, A.R. Metabolic profiling reveals altered nitrogen nutrient regimes have diverse effects on the metabolism of hydroponically-grown tomato (*Solanum lycopersicum*) plants. *J. Exp. Bot.* **56**, 309–321 (2005).
38. Broeckling, C.D. *et al.* Metabolic profiling of *Medicago truncatula* cell cultures reveals the effects of biotic and abiotic elicitors on metabolism. *J. Exp. Bot.* **56**, 323–336 (2005).
39. Schnee, C. *et al.* The products of a single maize sesquiterpene synthase form a volatile defense signal that attracts natural enemies of maize herbivores. *Proc. Natl. Acad. Sci. USA* **103**, 1129–1134 (2006).
40. Suzuki, H. *et al.* Methyl jasmonate and yeast elicitor induce differential transcriptional and metabolic re-programming in cell suspension cultures of the model legume *Medicago truncatula*. *Planta* **220**, 696–707 (2005).
41. Tohge, T. *et al.* Functional genomics by integrated analysis of metabolome and transcriptome of *Arabidopsis* plants over-expressing an MYB transcription factor. *Plant J.* **42**, 218–235 (2005).
42. Goossens, A. *et al.* A functional genomics approach toward the understanding of secondary metabolism in plant cells. *Proc. Natl. Acad. Sci. USA* **100**, 8595–8600 (2003).
43. Morikawa, T. *et al.* Cytochrome P450 *CYP710A* encodes the sterol C-22 desaturase in *Arabidopsis* and tomato. *Plant Cell* **18**, 1008–1022 (2006).
44. Hirai, M.Y. *et al.* Elucidation of gene-to-gene and metabolite-to-gene networks in *Arabidopsis* by integration of metabolomics and transcriptomics. *J. Biol. Chem.* **280**, 25590–25595 (2005).
45. Tagashira, N. *et al.* The metabolic profiles of transgenic cucumber lines vary with different chromosomal locations of the transgene. *Cell. Mol. Biol. Lett.* **10**, 697–710 (2005).
46. Schauer, N. *et al.* Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nat. Biotechnol.* **24**, 447–454 (2006).
47. Askenazi, M. *et al.* Integrating transcriptional and metabolite profiles to direct the engineering of lovastatin-producing fungal strains. *Nat. Biotechnol.* **21**, 150–156 (2003).
48. Urbanczyk-Wochniak, E. *et al.* Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep.* **4**, 989–993 (2003).
49. Jenkins, H. *et al.* A proposed framework for the description of plant metabolomics experiments and their results. *Nat. Biotechnol.* **22**, 1601–1606 (2004).
50. Scholz, M., Gatzek, S., Sterling, A., Fiehn, O. & Selbig, J. Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics* **20**, 2447–2454 (2004).
51. Scholz, M., Kaplan, F., Guy, C.L., Kopka, J. & Selbig, J. Non-linear PCA: a missing data approach. *Bioinformatics* **21**, 3887–3895 (2005).
52. Steuer, R., Kurths, J., Fiehn, O. & Weckwerth, W. Observing and interpreting correlations in metabolomic networks. *Bioinformatics* **19**, 1019–1026 (2003).
53. Urbanczyk-Wochniak, E. *et al.* Profiling of diurnal patterns of metabolite and transcript abundance in potato (*Solanum tuberosum*) leaves. *Planta* **221**, 891–903 (2005).
54. Ishizaki, K. *et al.* The critical role of *Arabidopsis* electron-transfer flavoprotein: Ubiquinone oxidoreductase during dark-induced starvation. *Plant Cell* **17**, 2587–2600 (2005).
55. Ishizaki, K. *et al.* The functional association between *Arabidopsis* electron transfer flavoprotein (ETF) and electron transfer flavoprotein ubiquinone oxidereductase (ETFQO) during dark induced starvation. *Plant J.* (in press).
56. Stitt, M. & Fernie, A.R. From measurements of metabolites to metabolomics: an 'on the fly' perspective illustrated by recent studies of carbon-nitrogen interactions. *Curr. Opin. Biotechnol.* **14**, 136–44 (2003).
57. Schad, M., Mungur, R., Fiehn, O. & Kehr, J. Metabolic profiling of laser microdissected vascular bundles of *Arabidopsis thaliana*. *Plant Methods* **1**, 2 (2005).
58. Roessner-Tunali, U. *et al.* Kinetics of labelling of organic and amino acids in potato tubers by gas chromatography-mass spectrometry following incubation in ^{13}C labelled isotopes. *Plant J.* **39**, 668–679 (2004).
59. Tieman, D. *et al.* Tomato aromatic amino acid decarboxylases participate in synthesis of the flavor volatiles 2-phenylethanol and 2-phenylacetaldehyde. *Proc. Natl. Acad. Sci. USA* **103**, 8287–8291 (2006).
60. Roessner-Tunali, U. *et al.* De novo amino acid biosynthesis in potato tubers is regulated by sucrose levels. *Plant Physiol.* **133**, 683–692 (2003).
61. Kopka, J. *et al.* GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* **21**, 1635–1638 (2005).
62. Schauer, N. *et al.* GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett.* **579**, 1332–1337 (2005).
63. Tolstikov, V.V. & Fiehn, O. Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Anal. Biochem.* **301**, 298–307 (2002).

2

Nonsupervised Construction and Application of Mass Spectral and Retention Time Index Libraries from Time-of-Flight Gas Chromatography–Mass Spectrometry Metabolite Profiles

Alexander Erban, Nicolas Schauer, Alisdair R. Fernie, and Joachim Kopka

Summary

Gas chromatography–mass spectrometry (GC–MS) is routinely applied to the metabolite profiling of biological samples. Time-of-flight (TOF)-GC–MS metabolite profiling is based on highly reproducible electron impact ionization. Single chromatograms may comprise 200–1000 mass spectral components. The nature and composition of these mass spectral components depend on the choice of metabolite extraction, type of biological sample, and experimental condition. The components represent mass spectral tags (MSTs) of volatile metabolites or metabolite derivatives. Identification of MSTs is the major challenge in GC–MS metabolite profiling. We describe methods suitable for the automated construction of mass spectral and retention time index databases from large sets of TOF-GC–MS profiles. Application of these libraries for automated identification by pure reference compounds and classification of hitherto unidentified MSTs from biological sources is demonstrated.

Key Words: Metabolite profiling; electron impact ionization; time-of-flight GC–MS; mass spectral matching; retention time index; metabolite classification.

1. Introduction

One of the major challenges in gas chromatography–mass spectrometry (GC–MS)-based metabolite profiling is the identification of the multitude of hitherto unidentified metabolic components from extracts of diverse biological samples ***(1,2)***. Automated deconvolution of single GC–MS chromatograms generates hundreds of mass spectral tags (MSTs) **(Fig. 1)**. MSTs were previously defined as mass spectra of metabolites or metabolite derivatives ***(3,4)***, which can be unambiguously identified by mass abundance or fragment composition and chromatographic retention behavior. As a rule of thumb, less than 30–40% of the detected MSTs can currently be linked to known metabolites. Unidentified MSTs are not necessarily artefacts of the GC–MS profiling technology. These MSTs can be shown to represent metabolites by in vivo labeling of organisms with stable isotopes, for example, labeling of microbial cultures by U-^{13}C-glucose ***(5)*** or feeding of $^{13}CO_2$ to photoautotrophic organisms. Thus, efforts to identify MSTs will be crucial for the further development and general applicability of GC–MS-based metabolite profiling ***(1–3)***.

Identification of MSTs is performed through two complementary approaches. The "top down" approach whereby metabolite identities are unravelled by taking, for example, a single MST of interest and establishing its structure through stepwise purification and complete structural elucidation. This approach is highly time-consuming. "Top down" identification is only recommended if the biological function of the unknown MST is clearly established and if the importance of the hitherto unknown metabolite justifies the task. The second, i.e., "bottom up," approach in which metabolites of interest to a particular researcher are analyzed by the purchase or synthesis of authentic standards is certainly less time demanding and, thus, appears to be more efficient. Identification is easily performed through standard addition experiments of pure reference compounds ***(6–9)***. Both mass spectral matching and co-chromatography can routinely be established in different laboratories and can thus meet the general prerequisites of unambiguous chemical identification ***(3)***. In summary, metabolite identification can be repeated with all GC–MS equipment and is easy to cross-validate between many laboratories across the world.

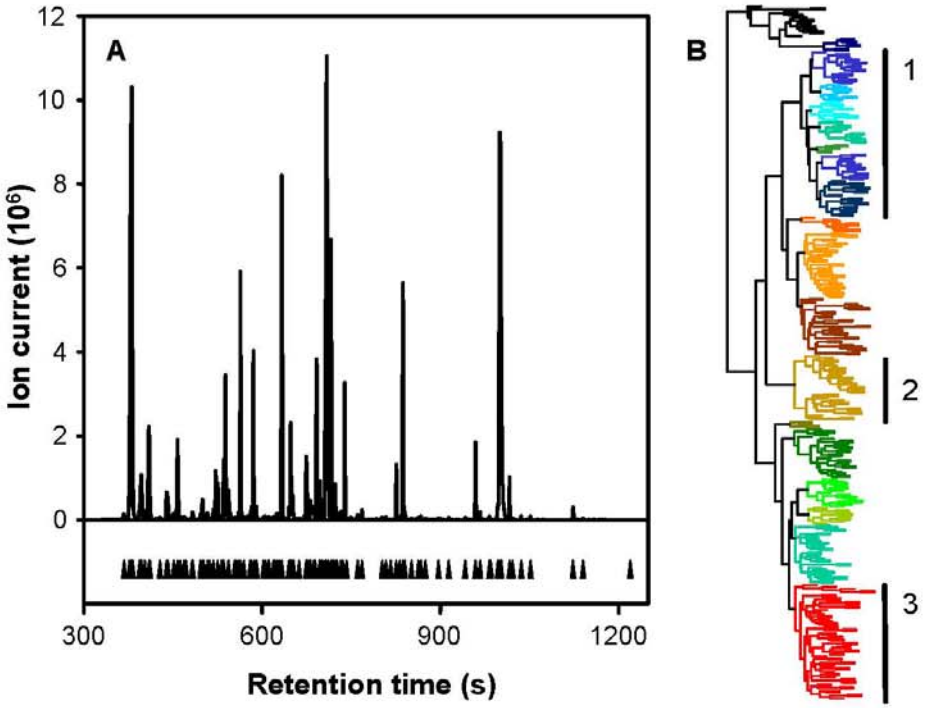


Fig. 1. Metabolite profile of an intercellular yeast extract **(A)**. Tics below the chromatogram trace indicate positions of automated deconvolutions of mass spectra. Mass spectral components obtained from yeast metabolite profiles were clustered with mass spectra of pure reference metabolites **(B)**. Major clusters represent (1) sugars, polyols, and polyhydroxy acids, (2) phosphorylated compounds, and (3) amino acids.

We describe a largely automated method for the highly reproducible generation of MSTs and mass spectral/retention time index (MSRI) libraries. The method is designed to suppress artifacts of chemical derivatization by timed and automated in-line derivatization of metabolic extracts. Furthermore, the high degree of automation supports increased reproducibility of retention time behavior, as determined by Kovàts' retention time indices (RI) ***(10)***. In parallel, mass spectral characteristics are quality controlled by in-build auto-tuning routines of the GC-time-of-flight (TOF)-MS system ***(11)***.

The availability of curated MSRI libraries as well as nonsupervised MSRI libraries, i.e., automated generation of MST compendia from well characterized and defined biological samples, facilitates identification of metabolites in diverse biological samples ***(3,12)*** and integrates use of commercially available mass spectral libraries ***(13,14)***, which lack retention time characteristics. The aim of this method description is to enable mass spectral library searches with single bait mass spectra of a reference substance that allow clear identification by mass spectral match and RI **(Fig. 2)**. Moreover, the hit lists of these mass spectral searches are utilized to discover candidate component MSTs of highly similar chemical nature as compared with the bait and, thus, facilitate classification of as yet unidentified MSTs **(Fig. 3)**.

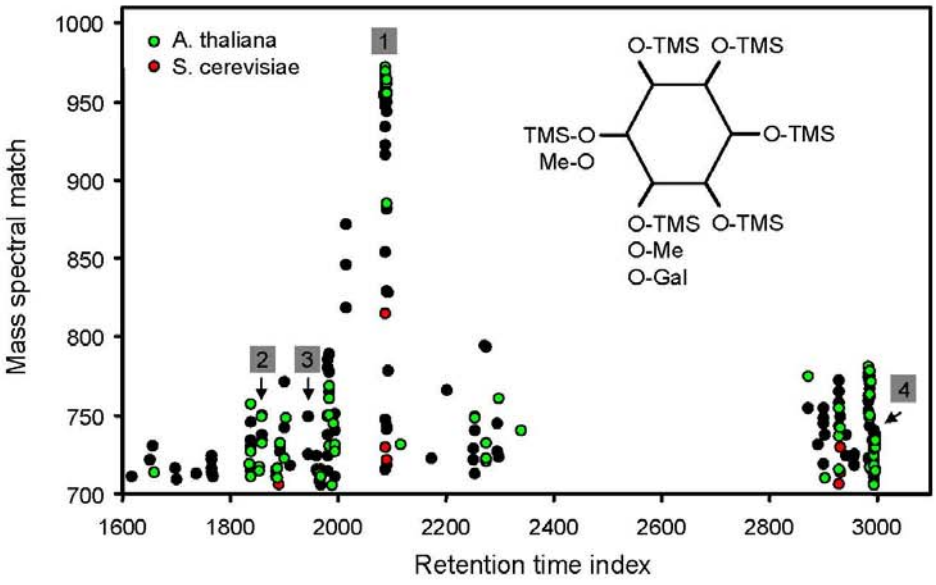


Fig. 2. Mass spectral hit list of *myo*-inositol (6TMS). Identified mass spectral tags were (1) *myo*-inositol (6TMS), (2) 3-*O*-methyl-D-*chiro*-inositol (5TMS), (3) 4-*O*-methyl-*myo*-inositol (5TMS), and (4) α-D-galactopyranose-(1,3)-*myo*-inositol (9TMS). Occurrence in *Arabidopsis thaliana* and *Saccharomyces cerevisiae* is color coded. Mass spectral matching was performed without limits or constraints.

2. Materials

2.1. Sampling and Metabolite Extraction

1. Methanol gradient grade for liquid chromatography (Merck, Darmstadt, Germany; cat. no. CAS 67-56-1).
2. Chloroform for liquid chromatography (Merck; cat. no. CAS 67-66-3).
3. Bidistilled water approx 0.055 S/cm (USF Deutschland GmbHRansbach-Baumbach, Germany; cat. no. USF 800).
4. Ribitol (Sigma, Munich, Germany; cat. no. CAS 488-81-3).
5. DL-Alanine, 2,3,3,3-d_4 (Sigma; cat. no. CAS 53795-92-9).
6. D(–)-Isoascorbic acid (Sigma; cat. no. CAS 89-65-6).
7. Methyl nonadecanoate (Sigma; cat. no. CAS 1731-94-8).
8. 1.5-mL Safe-lock, tapered-bottom microvial (Eppendorf, Hamburg, Germany).
9. 2.0-mL Safe-lock, round-bottom microvial (Eppendorf).
10. Microcentrifuge 5417 (Eppendorf).
11. Oscillating ball mill MM200 (Retsch GmbH and Co. KG, Haan, Germany).
12. Teflon adaptor for 1.5- to 2.0-mL microvials (Retsch GmbH and Co. KG).
13. VA 5-mm steel balls (Th. Geyer Berlin GmbH, Berlin, Germany).
14. VR Maxi standalone vacuum concentrator with rotors R96-13 and R120-111 (Jouan Nordic, Allerod, Denmark).
15. HBP hold-back vacuum pump (Ilmvac GmbH, Ilmenau, Germany).
16. Polystat K6-1 cycling thermostat (P. Huber GmbH, Offenburg, Germany).
17. 15- and 50-mL plastic tubes with screw caps (Falcon™ Conical Centrifuge Tubes, BD Biosciences, San Jose, CA).
18. Orange silica gel (Carl Roth GmbH, Karlsruhe, Germany; cat. no. 77.1).
19. Argon 5.0 (Messer-Griesheim GmbH, Krefeld, Germany).

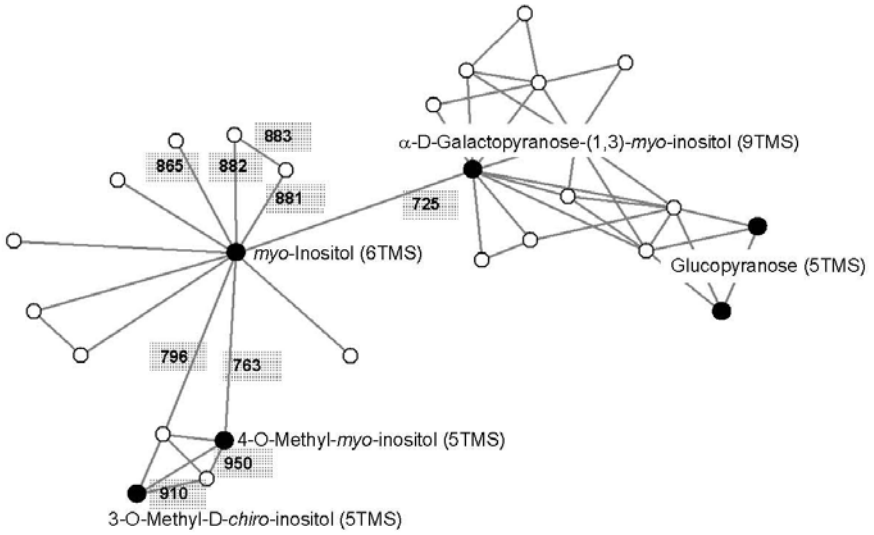


Fig. 3. Proximity map of a search for mass spectral similarity among identified and unidentified mass spectral tags from GC–MS profiles of biological sources. The search was initiated at *myo*-inositol (6TMS). Open circles represent hitherto unclassified mass spectra, and connecting edges represent best mass spectral match as partially indicated in shaded boxes. Mass spectral matching was performed in the mass range, *m/z* 85–600, with minimum abundance set to 50.

2.2. Chemical Derivatization

1. CTC Combi PAL autosampler and PAL cycle composer software v1.5.0 (CTC Analytics AG, Zwingen, Switzerland). The chosen configuration comprises an agitator–incubator oven, a 98-sample tray for 2.0-mL vials, a 32-sample tray for 10- to 20-mL vials, three 100-mL solvent reservoirs, i.e., a syringe wash station, and a liquid version 25-µL syringe kit mounted to the robotic autosampler arm.
2. Methoxyamination reagent: methoxyamine hydrochloride (Sigma; cat. no. CAS 593-56-6) is dissolved at 20 mg/mL in pure pyridine (Merck; cat. no. CAS 110-86-1). This reagent is prepared immediately before analysis in 1-mL aliquots and loaded into the first reagent reservoir of the CTC Combi PAL autosampler (*see* **Note 1**).
3. Per-silylation reagent: 1-mL vials of *N*-methyl-*N*-(trimethylsilyl)trifluoroacetamide (MSTFA; Macherey and Nagel, Düren, Germany; cat. no. CAS 24589-78-4) is loaded into the second reagent reservoir (*see* **Note 1**).
4. Solvents for syringe washes were *n*-hexane (Fisher-Scientific GmbH, Schwerte, Germany; cat. no. CAS 110-54-3) and ethylacetate (Merck; cat. no. CAS 141-78-6).
5. RI standard mixture: *n*-alkanes are dissolved in pyridine (Merck; cat. no. CAS 110-86-1) at a final concentration of 0.22 mg/mL each and loaded into the agitator–incubator oven of the CTC Combi PAL autosampler (*see* **Note 2**). The following substances are combined: *n*-decane (RI 1000; cat. no. CAS 124-18-5), *n*-dodecane (RI 1200; cat. no. CAS 112-40-3), *n*-pentadecane (RI 1500; cat. no. CAS 629-62-9), *n*-octadecane (RI 1800; cat. no. CAS 593-45-3), *n*-nonadecane (RI 1900; cat. no. CAS 629-92-5), *n*-docosane (RI 2200; cat. no. CAS 629-97-0), *n*-octacosane (RI 2800; cat. no. CAS 630-02-4), *n*-

dotriacontane (RI 3200; cat. no. CAS 544-85-4), and *n*-hexatriacontane (RI 3600; cat. no. CAS 630-06-8). All substances were obtained from Sigma.
6. 1.1 CTVG crimp-cap vial (Chromacol, Trumbull, CT).
7. R11-Sil-r/w magnetic crimp cap (CS-Chromatography Service GmbH, Langerwehe, Germany).
8. Adjustable 11-mm crimp-cap sealer (Supelco, Munich, Germany).

2.3. TOF-GC–MS

1. Pegasus III TOF mass spectrometer (LECO Instrumente GmbH, Mönchengladbach, Germany).
2. Agilent 6890N gas chromatograph, split/splitless injector with electronic pressure control up to 150 psi (Agilent, Böblingen, Germany).
3. Conical single taper split/less liner with glass wool (Agilent), deactivation reagent (DMDCS, Restek GmbH, Bad Homburg, Germany), toluene (Sigma; cat. no. CAS 108-88-3), methanol (Merck; cat. no. CAS 67-56-1) (*see* **Note 3**), 7-mL glass tubes (cat. no. 23 175 11 59) with screw caps (cat. no. 29 990 12 04) (Schott, Mainz, Germany).
4. RTX-5Sil MS capillary column, 30-m length, 0.25-mm inner diameter, 0.25-µm film thickness, and a 10-m Integraguard precolumn (Restek GmbH).
5. Helium 5.0 carrier gas (Air Liquide, Magdeburg, Germany).

2.4. GC–MS Data Processing

1. ChromaTOF chromatography processing and mass spectral deconvolution software, v1.00, driver 1.61 (LECO Instrumente GmbH).
2. Automated mass spectral deconvolution and identification system AMDIS (National Institute of Standards and Technology [NIST], Gaithersburg, MD).
3. NIST mass spectral search and comparison software v2.0 (NIST).
4. Microsoft Office Word 2003 (Microsoft Corporation), Excel 2003 (Microsoft Corporation), software package for exploratory data analysis and statistical modelling, S-Plus 2000 standard edition release 3 (Insightful, Berlin, Germany).

3. Methods

Metabolite turnover is extremely rapid as compared to mRNA or protein turnover. Analysis of metabolite composition and changes in pool sizes, therefore, requires fast and reproducible metabolic inactivation. Samples are best shockfrozen and kept below –60°C until extraction. Maintenance of metabolic inactivation during extraction and workup procedures is essential for a robust and repeatable representation of *in situ* metabolite composition ***(15)***. We describe two exemplary protocols of metabolite extraction from plant material and liquid microbial cultures. We are fully aware that the choice of metabolic inactivation and extraction protocol may influence and indeed determine the scope of metabolites that can be monitored by subsequent metabolite profiling. Variations of extraction protocols may be pursued to broaden the spectrum of metabolites that are accessible to metabolite profiling (*see* **Note 4**) or to perform integrated analyses of metabolome, proteome, and transcriptome ***(16)*** (*see* Chapter 5).

The essence of metabolite profiling is discovery of novel marker metabolites and determination of relative changes of metabolite pool sizes in comparison to reference samples ***(1,2)***. This approach necessitates thorough control experiments, monitoring of GC–MS system performance, and check of laboratory contaminations, which may arise from solvent and reagent impurities or leakage of vial and septum material. For these reasons, nonsample control experiments are indispensable. All chemicals and containers need to be of highest available purity. Please consider that autoclaved material, although sterile, may nevertheless be chemically contaminated.

3.1. Experimental Design and Preparation of Samples for Metabolite Profiling

Make sure to include a set of nontreated control samples in each experiment and analysis. For a large series of analyses prepare and store a large batch of reference material. Take an additional set of

control samples from this batch of reference material for each subset of analyses. Results from this reference material allow the experimenter to control for day-to-day and week-to-week variability. Thus, discovery of marker metabolites and relative changes in metabolite levels can be distinguished from accidental contaminations or changes in instrumental sensitivity.

Provide at least 6 (better 8–16) replicate samples of each experimental condition. Perform replications at the level of individual plants or cell cultures rather than repeating assays of the same sample ***(6–9,15)***. Pooling of samples from a set of plants or cell cultures and repeated analyses of this pool is advised when sample size is small. Analysis of sample pools, however, is less informative with respect to the underlying variability inherent to the experiment and nature of biological samples.

3.1.1. Metabolic Inactivation and Extraction of Plant Material

1. Shock-freeze plant material in liquid nitrogen and keep below –60°C throughout processing. Use precooled 2.0-mL safe-lock micro vials or wrap samples in precooled aluminium foil. Store either in liquid nitrogen or at –80°C until further processing. The amount of required sample may vary depending on species and plant organ. Always perform test analyses when analyzing previously unknown samples or novel experimental condition. The following protocol describes a typical analysis that is optimized for 60-mg fresh weight (±5–10%) of dicot leaves. Monocot leaves or root material may require more material (factor 2–4), whereas storage organs, flowers, or cold-stressed material may be performed with smaller amounts (factor 0.1–0.5). The optimum sample load is best determined by adjusting the major metabolic components to the upper detection limit of GC–MS, while still avoiding peak overload (*see* **Subheading 3.4.**).
2. The preparation of representative aliquots from large samples, greater than 125 mg (fresh weight), requires homogenization using a precooled mortar and pestle and subsequent generation of small aliquots of the desired amount of material. Keep samples in liquid nitrogen throughout the process. Avoid condensing ice and be careful not to spill the final powder by boiling liquid nitrogen. The powder may be stored in liquid nitrogen or in –80°C freezer using screw-cap or safelock vials. Be careful to evaporate residual liquid nitrogen at –80°C before caps are sealed.
3. Small samples, 5–125 mg (fresh weight), are homogenized using steel balls that fit into 2.0-mL round-bottom microvials. Sets of 5–10 microvials are mounted onto an oscillating ball mill and exposed to two 3-min bursts at 15/s frequency. Steel balls, microvials, and mounting adaptor need to be precooled in liquid nitrogen. Homogenized samples are extracted within microvials without removal of the steel balls. Sample weight is best determined after shock-freezing. Differential weighing of frozen powder or nonhomogenized material can be performed in cooled 2.0-mL microvials before adding steel balls. Avoid high air humidity and use dry ice for cooling to obtain stable zero point calibration.
4. Take frozen 2.0-mL microvials with homogenized samples from the freezer and add 360 µL of extraction mixture (*see* **Notes 5** and **6**). The extraction mixture needs to be precooled to –20°C and is best degassed by bubbling argon or nitrogen gas. Use an oil filter between gas supply and high-performance liquid chromatography bubbling device. Shake samples thoroughly using a vortex mixer and keep on ice until all samples are processed.
5. Shake all samples simultaneously 15 min at 70°C and subsequently cool to room temperature. Solvent evaporation may generate excess pressure. Vent microvials after 1 min incubation at 70°C and reclose vials thoroughly.
6. Add 200 µL $CHCl_3$, shake thoroughly using a vortex mixer, and incubate at 37°C.
7. Add 400 µL H_2O to induce phase separation, shake thoroughly using a vortex mixer, and separate liquid and solid phases in a microcentrifuge for 5 min at approx 22,000 g. Addition of H_2O may be omitted for a joined analysis of the lipophilic and polar metabolic complement of the sample.
8. Take a 10-µL aliquot of the upper phase, which contains the polar metabolic complement of the sample, and transfer into a crimp cap-tapered glass vial suitable for GC–MS analyses (*see* **Note 7**). In the following, we describe automated analysis of 10 µL of the polar or a combined liquid extract. In case of manual processing, 1.5–2.0 mL safe-lock microvials may be used to dry, transport, and store metabolic extracts (*see* **Subheading 3.2.**). For analysis of the lipophilic metabolic complement, take a 100-µL aliquot of the lower liquid phase and process manually. The analysis of the lipophilic complement induces strong chromatographic memory effects and is not recommended for high-throughput split or splitless GC–MS injection.
9. Dry 10 µL samples in a vacuum concentrator for a minimum of 2 h at room temperature or lyophilize larger sample volumes over night.

3.1.2. Metabolic Inactivation and Extraction of Yeast Liquid Cultures

The major challenge in metabolite profiling of microbial cultures is the separation of intracellular metabolites from secreted metabolites and residual components of liquid growth media, the so-called footprint, while rapidly inactivating metabolism during sampling. Typically, cell suspensions are rapidly sprayed into precooled polar organic solvents, such as methanol, which dilute the media and shock-freeze the cells ***(17–19)***. We recommend growth media spiked with nonmetabolized low molecular weight compounds for the control of residual liquid medium, which is unavoidably trapped in the cellular periplasm. In the case of yeast we successfully used lactose, which cannot be utilized by yeast, at 1–10% (w/w) concentration of the major carbon source in the growth medium. We furthermore suggest use of synthetic-defined growth media (SD) instead of complex media. Complex media contains numerous compounds in high concentrations. These substances will obscure intracellular metabolites even in cases of only small medium contaminations.

1. Prepare 5-mL yeast batch cultures in SD medium and time the sampling to the late logarithmic or to stationary growth phase (OD_{595} = 1.8). Follow general recommendations for yeast growth ***(17–19)***. Make sure to prepare noninocculated samples for nonsample control of the experiments. Avoid unwanted chemical contaminations of the liquid cultures. Sterilized glassware and media are devoid of microbial contaminations but might nevertheless have received chemical deposits from the autoclave. Media components may decompose while exposed to high temperatures.
2. Sample the complete culture at routine growth temperature, 28°C, by rapid decanting or use temperature equilibrated disposable pipet tips for sampling 5-mL aliquots from larger batches. Avoid slow temperature changes before sampling. Continue to agitate batch cultures until sampling. Thus, sedimentation of cells and changes in mechanical stress are circumvented.
3. Rapidly mix 5 mL medium with 20 mL precooled 60% methanol, methanol:water, 6:4 (v/v). 60% methanol is best prepared as a large batch and partitioned into 50-mL screw-cap plastic tubes, which are kept before and after sampling in a methanol/dry ice bath at approx –60°C.
4. Spin down cells no longer than 5 min at approx 3200 g in a temperature-controlled centrifuge preset to –20°C.
5. Immediately after centrifugation, collect plastic tubes into the methanol/dry ice bath. Decant supernatant cautiously and perform an optional gentle rinse with a small volume of precooled 60% methanol. During temperature adjustment the supernatant might get slightly turbid but should not freeze solid.

The following steps can be downscaled according to the initial concentration of cells in suspension culture as determined by OD_{595} of diluted samples. The following volumes are as required for a 5-mL culture of OD_{595} = 1.8.

6. Add 374 µL extraction mixture for yeast intercellular metabolites immediately (*see* **Note 8**). The extraction mixture needs to be precooled to –20°C and is best degassed (*see* **Steps 4** and **5**). At this step the cells should easily resuspend. If cells form a semi-solid viscous pellet, the temperature control was inadequate for metabolite profiles and needs to be optimized. Critical steps are centrifugation and time between decanting of supernatant and resuspension into extraction mixture. Slightly viscous yeast pellets may be resuspended in small droplets of icecold water prior to adding the extraction mixture. Metabolite profiling of these samples is not recommended.
7. Transfer resuspended samples from 50-mL plastic tubes into 7-mL screw-cap glass tubes for simultaneous extraction, 15 min at 70°C. Shake glass tubes intermittently and depressurize at least once. Allow to cool for 5 min at room temperature.
8. Add 188 µL $CHCl_3$ and extract 10 min at 30°C with intermittent vigorous shaking using a vortex mixer.
9. Add 75 µL of bidistilled H_2O, spin down cellular debris, and transfer a 10-µL aliquot of the combined polar and lipophilic extract into a crimp cap-tapered glass vial suitable for GC–MS analyses. Phase separation into a polar and lipophilic metabolic complement may be induced by adding 400 µL H_2O prior to centrifugation. Subsequent steps are as previously described (*see* **Subheading 3.1.1.**).

3.2. Storage and Transport of Metabolite Extracts

Metabolite extracts are best stored at low temperatures and under nonoxidizing conditions. If possible, long periods of storage and transport should be avoided. Samples can be transported and stored for up to 4 wk. Longer periods have not been tested.

1. After drying samples in a vacuum concentrator or lyophilization, flush the vacuum system with an inert gas, such as argon or nitrogen, instead of ambient air before removing samples.

2. Seal GC vials under inert gas using magnetic crimp caps and an adjustable crimpcap sealer. Seal vials in plastic bags with silica gel. Combine the full number of vials comprising one experiment in single bags.
3. Transport sealed bags for short periods at room temperature otherwise on dry ice and store at –20 or –80°C.
4. Allow temperature equilibration at room temperature before opening bags for further analysis.

3.3. TOF-GC–MS Metabolite Profiling

Profiling of metabolite extracts involves a two-step chemical derivatization, which (1) substitutes carbonyl moieties through methoxyamination and (2) comprises a per-silylation prior to the GC–MS analysis of the reaction products ***(6–9)***. Samples are injected while dissolved in silylation reagent. Major sources of analytical variability are the imprecise dispensing of reagent volumes and the variable timing of the per-silylation reaction. In typical experiments, 50–100 samples are processed. Chemical derivatization was hitherto performed simultaneously on a batch of samples prior to injection. Thus the exposure time to the silylation reagent of first and last sample within a batch differed considerably, i.e., 50–100 h in set-ups of 60 min per single GC–MS run. As a result, instable derivatives discomposed, side products of silylation reagents accumulated, and slow evaporation caused notable sample concentration. An optimization of the chemical reaction and GC–MS analysis was, therefore, in high demand.

We employ a CTC Combi PAL with a single syringe autosampler for automated and timed in-line derivatization, vial transport, and injection for GC–MS analysis. Vials are transported from the vial tray to positions within the agitator–incubator oven and finally back to the injection position by means of magnetic crimp caps. In short, in-line chemical derivatization requires samples to be dried within GC glass vials and sealed under nitrogen or argon. Each sample is processed in four equal time intervals of 45 min each. The first two intervals are assigned to methoxyamination (90 min), the third to per-silylation (45 min), and the fourth to a single slow or alternately two fast GC–MS runs per sample (total time <45 min). A typical GC-TOF-MS profile of a preparation of intracellular yeast metabolites is shown in **Fig. 1**.

3.3.1. In-Line Chemical Derivatization

1. The following instructions require 10 µL of metabolic extracts to be dried in 1.1 CTVG crimp-cap vials. The sealed vials are positioned on the sample tray and kept at ambient temperature (*see* **Note 9**).
2. Methoxyamination: the first vial is moved to position 1 of the agitator–incubator oven, which is set to constant 40°C. A 10-µL volume of methoxyamination reagent is dispensed into the vial. The vial is then agitated twice for 45 min.
3. Per-silylation: after 90 min, agitation is interrupted by dispensing 17.5 µL persilylation reagent. Then 2.5 µL of a retention time standard mixture are added. Agitation is resumed for an additional interval of 45 min at 40°C.
4. At the end of the last interval the GC vial is moved back to the initial position on the sample tray and 1 µL is injected for GC–MS analysis (*see* **Note 10**). Processed vials are kept on the sample tray until discarded.
5. For automated high-throughput analysis, samples are processed in parallel with a time lag of 45 min each. Four positions of the agitator–incubator oven are used, three for derivatization of samples and one to store the retention time index standard mixture of *n*-alkanes (*see* **Note 2**). The most recent sample in the process is always subject to the first methoxyamination interval. Prior samples are in the second methoxyamination period, the per-silylation interval or in the process of GC–MS analysis, respectively.
6. Syringe washes are performed between all dispensing procedures (*see* **Note 11**).
7. Automation using the Combi PAL autosampler can be performed with three basic programming parts. The first part primes the in-line derivatization process and ends with injecting the first sample, while the following two samples are already under derivatization. The second part comprises three methods that allow an "endless"cycle, each cycle ending with an injection. The final programming part contains methods that end in-line derivatization by processing the last samples of an analysis series and then safely shuts down the system.

3.3.2. TOF-GC–MS

1. Injection parameters: injection of a 1-µL sample is performed at 230°C in splitless mode with helium carrier gas flow set to 0.6 mL/min. Purge time is 1 min at 20 mL/min flow. The flow rate is kept constant with electronic pressure control enabled. Optionally and especially recommended in cases of high metabolite concentrations, injection is performed in split mode with the split ratio adjusted 1:25. As a rule of thumb, split injection may be prone to discrimination of high-boiling metabolic components, whereas splitless injection may, in rare cases, result in peak shape artifacts for low-boiling components. These artifacts occur in few chromatograms and result in different degrees of peak splitting and shoulder formation. For suppression of this peak shape, artifact either injection at decreased flow or a 2-min pressure pulse at 110 psi during injection. However, a robust suppression of this artifact for all biological samples can currently not be recommended.
2. Chromatography parameters: chromatography is performed using a 30-m RTX- 5Sil MS capillary column with an integrated guard column. The temperature program starts in isothermal mode set to 1 min at 70°C. The isothermal step is followed by a 9°C/min ramp to 350°C. The final temperature is kept constant for 5 min. Cooling is performed as fast as instrument specifications allow. The transfer line temperature is set to 250°C and matches ion source conditions.
3. Mass spectrometer parameters: the ion source is set to maximum instrument specifications, 250°C. High-boiling metabolic components exhibit increased peak tailing at lower temperature settings. The recorded mass range is m/z = 70–600 at 20 scans/s. Mass spectrometric solvent delay with filaments turned off is 6.6–7.5 min, the remaining chromatography including cool down periods is fully monitored. Manual mass defect is set to 0, filament bias current is –70 V, and detector voltage is approx 1700–1850 V depending on detector age. The instrument tune is automated and performed without EPA tune compliance.

3.4. Automated Deconvolution of Mass Spectra

Automated deconvolution of MSTs from GC–MS metabolite profiles is crucial for increased accuracy of metabolite identification and detection (**Fig. 1**). Deconvolution is the process of locating MSTs, also called mass spectral components, in GC–MS chromatograms and the subsequent automated purification of the mass spectral scans at peak apex from electronic and chemical background noise and cross-contaminating fragments of coeluting compounds. Both the ChromaTOF software of LECO GC-TOF-MS systems and the technology platform-independent automated mass spectral deconvolution and identification system, AMDIS, may be used to this purpose ***(3,13–14)***. When using AMDIS, files are best exported in CDF file format after baseline correction within the ChromaTOF software. Large GC-TOF-MS files, such as those with fast scanning acquisition rates, may be impossible to load into AMDIS using standard desktop computers. Here, we describe the use of the ChromaTOF software for automated deconvolution and construction of MSRI libraries. MSRI libraries may contain either manually curated and selected entries of identified compounds or have the purpose to provide full automatically generated collections of mass spectra from single or multiple GC-TOF-MS profiles. This process we would like to term nonsupervised construction.

1. Chromatograms are processed by ChromaTOF software with activated baseline tracking and offset set to "just above noise," smoothing and peak width are set to three and six, respectively. The signal-to-noise threshold is set to minimum 2.0 and the number of deconvolutions is unlimited.
2. RIs are generated for each individual chromatogram in two steps: first a mass spectral library search is conducted to identify all expected *n*-alkanes in each chromatogram, then, retention times of the *n*-alkanes are used for chromatogram specific RI calculation. The mass spectral library search for *n*-alkane identification is restricted to the mass range m/z = 80–600 and threshold abundance set to 20, i.e., 2% of the base peak intensity. Further criteria for identification are peak height and area in total ion chromatography mode (TIC), as well as occurrence of respective molecular ions for each *n*-alkane. The retention times of the expected and verified *n*-alkanes are transferred into a chromatogram specific retention index method and the chromatogram subsequently processed with the same settings. Overloaded peaks must be avoided or excluded in order to maintain high RI accuracy.
3. Chromatogram processing results are exported to text files. All available information for each deconvoluted peak or MST is exported including auxiliary information, such as retention time index, retention time, unique mass, total signal to noise, and full mass spectrum in absolute intensity format.
4. These text files can be imported and modified in Microsoft Excel and Word. More efficient is a customized automated programmed conversion into the MSP format for import into NIST02 and AMDIS

software, which needs to add RI information for the generation of MSRI libraries. During this process auxiliary information, such as user comments, can be tagged as synonyms. MSTs can be removed or selected by signal to noise, peak purity, peak width, or RI thresholds. Thus, data can be specifically selected for import into NIST02 software and information can be added. A typical example of an identified mass spectrum is shown in the following:

Name: EITTMS_163001-101_METB_1627.14_L-Glutamic acid (3TMS)
Synon: SOURCE_CHROMATOGRAM:1185EK12_1627.1
Synon: NAME:L-Glutamic acid (3TMS)
Synon: MATCH:[834; L-Glutamic acid (3TMS)]
Synon: MPIMP-ID:163001-101-1
Synon: QM:246|363|128|348|156
Synon: ROLE:METB
Synon: METABOLITE:DL-Glutamic acid
Synon: KEGG:C00025|C00302|C00217
Synon: TECHNOLOGY:GC-TOF-MS (EITTMS)|GC [M1]
Synon: RI:1627.1
Synon: RT:10.253 min
Synon: SP:Standard| Sigma| G-1251
Synon: DATE:2001.06.01
Comments: Kopka J, Max Planck Institute of Molecular Plant Physiology, Am Muehlenberg 1, D-14476 Golm, Germany
Formula: C14H33NO4Si3
MW: 363
CAS no.: 15985-07-6
DB no.: 799
Num Peaks: 151
70 11; 71 6; 72 38; 73 999; 74 102;
75 454; 76 33; 77 79; 78 6; 79 7;
80 2; 81 1; 82 5; 83 6; 84 164; ...

5. Chromatogram processing results can also be exported directly from the peak table of the ChromaTOF software to NIST02 without the need of programming skills. Deconvoluted MSTs can be either added to NIST02 user libraries or exported as MSP files. Customization of library entries within ChromaTOF software before export is highly restricted. However, NIST02 offers a full toolbox for editing mass spectral information. Thus, mass spectral libraries of manually selected, identified, and curated MSTs can be easily generated and maintained with the tools and options provided by NIST02 and ChromaTOF software.
6. Examples of annotated MSRI libraries comprising identified compounds as well as unidentified MSTs and MSRI libraries, which were fully generated in the nonsupervised mode, may be found at CSB.DB (http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html).

3.5. Comparison and Classification of Mass Spectra

The NIST02 mass spectral search and comparison software represents the most widely accepted standard tool for analysis of mass spectra generated by GC–MS systems ***(13–14)***. Systems' manufacturers optimize automated MS tuning with the aim to produce comparable mass spectra. NIST02 is mature in automation, algorithm, as well as user friendliness. However, the great challenge of identifying or at least classifying all hitherto unidentified metabolic components from GC–MS profiles of biological samples requires additional features that are not provided by NIST02. One of the most useful additional features for mass spectral comparisons is the integration of retention time index information into mass spectral comparisons. Only information on chromatographic retention will allow unambiguous identification of those stereo- and conformational isomers which cannot be distinguished by mass spectral criteria alone ***(3)***. In addition, mass spectral classification needs to be reconsidered for those MSTs that can not immediately be linked to a known metabolite. Here, we demonstrate first attempts to systematically deal with the challenge of identifying multiple unknown mass spectral tags from GC-TOF-MS profiles. Our present analyses are all performed using a single technology platform and a set of chromatograms that were produced on identical GC capillary columns. Transfer of our results to other technology platform appears to be feasible but still awaits thorough investigations.

3.5.1. Clustering

Mass spectra can be directly clustered using hierarchical clustering of Euclidian distance or any other algorithm of commercially or publicly available software packages for statistical analysis. An

alternative approach is clustering based on the generally accepted matching value generated by NIST02 mass spectral comparison software instead of Euclidian or other statistical distances. For "nonstandard" mass spectral distance measures and queries that incorporate RI information refer to our web pages, http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html ***(12)***. For the purpose of clustering a full matrix of pairwise similarity, measures of all MSTs and identified mass spectra needs to be defined through automated comparison and data export using NIST ***(3)***. We performed a combined analysis of identified and all MSTs that occur in yeast metabolite profiles (**Fig. 1A**). Clustering was performed as described using the S-Plus 2000 standard edition statistical software package. Clustering demonstrated the presence of major metabolite classes in GC-TOF-MS profiles, such as carbohydrates, amino acids, and organic phosphates (**Fig. 1B**). The mass spectrum of *myo*-inositol (6TMS), which we subsequently use as a test case, classifies to the sugar cluster. Most of the hitherto nonclassified MSTs sorted into clusters of identified metabolites. Thus, simple hierarchical clustering provides means to link unidentified MSTs to major metabolite classes. Some major clusters formed clear subdivisions. For example the carbohydrate cluster had disaccharide, monosaccharide, noncyclic polyol, and polyhydroxy carbonic acid branches. In total, up to 18 clear minor mass spectral clusters were found. However, clustering might lack resolution within the terminal branches of hierarchical trees.

3.5.2. Visualization of MSRI Search Results

For the resolution of mass spectral similarity at the level of single mass spectra, the NIST02 hit lists are unsurpassed, but are lacking in visualization. The additional RI information is best shown in bi-plots with axes of RI and mass spectral match (**Fig. 2**). These plots easily accommodate auxiliary information, for example, on occurrence of MSTs in different sample types and frequency of occurrence in cases of redundant mass spectral libraries, such as nonsupervised MSRI libraries from GC–MS profiles. These visualizations allow discovery of MSTs that exhibit similarity to the bait mass spectrum. In addition, structural similarities of identified mass spectra that become apparent as mass spectral similarity can be accessed. In our test case *myo*-inositol (6TMS) had among the top scoring identified mass spectra, methyl substituted inositols, ononitol (5TMS) (4-*O*-methyl-*myo*-inositol), pinitol (5TMS) (3-*O*-methyl-D-*chiro*-inositol), and an inositol conjugate, galactinol (9TMS) (α-D-galacto-pyranose-[1,3]-*myo*-inositol). Conformational isomers of *myo*-inositol, such as *chiro*- or *scyllo*-inositol, rank highest but are not jet included in this and the subsequent analysis.

3.5.3. Generation of Mass Spectral Proximity Maps

Hit lists of single MSTs present good means of discovery of best matching mass spectra but do not convey an overview of similarities between many MSTs. For this purpose proximity maps are best suited (**Fig. 3**). Proximity maps visualize the journey through the "space" of mass spectral matches present within a MSRI library. The process of generating a proximity map can be manually performed by starting a mass spectral search with a mass spectrum of interest, such as *myo*-inositol (6TMS). The aim of this process is to discover groups of related compounds based on mass spectral similarity. The initial hit list will contain redundant mass spectra of *myo*-inositol (6TMS) and one best hit, which as judged by RI or already known identity, represents a different compound. We travel to this compound along the best match (865 in **Fig. 3**) and will not use this connection in the same direction again throughout the remaining journey. Instead, we perform a mass spectral search with the found best hit. In our test case, the best match of this second search was *myo*-inositol (6TMS). Thus, we return to *myo*-inositol (6TMS) and close the connection in the reverse direction as well. We then continue with the next best match of *myo*-inositol (6TMS) (882 in **Fig. 3**). The proximity map is subsequently generated using the same rules. The journey can be terminated after a limited number of steps, a number of visited mass spectra, or at a threshold match value. Visualization of a proximity map can performed using network visualization tools such as Pajek software ***(20)***. The resulting map clearly shows that *myo*-inositol (6TMS) has a set of 11 directly linked MSTs in our present MSRI library within a similarity range of 725 to 865. Among those we found a set of four MSTs with high "internal" similarity (910–950), which represent two methyl substituted inositols and two putative still unidentified other methylinositols. Furthermore, we found an inositol conjugate, galactinol (9TMS), and a group of highly similar (match values not shown) MSTs, which form connections to

glucopyranoses that are highly similar in structure to the second conjugation partner, i.e., galactopyranose, of galactinol (9TMS).

4. Notes

1. Reagents are stored in 1-mL crimp-cap sample glass vials. These vials contain excess reagent but are replaced after 24 h in order to avoid aging and accumulation of contaminations.
2. The retention time index standard mixture contains high molecular weight *n*-alkanes, which tend to precipitate at low ambient temperature. The *n*-alkane mixture is best prepared at elevated temperature and during use is kept at 40°C within the heated agitator.
3. Deactivation of the glass insert liners reduces the number of cleaning cycles, which are required after liner exchange and increases column lifetime. For glass liner deactivation, dissolve 20 mL of DMDCS in 400 mL toluene and treat liners for 15 min in this solution. Then rinse twice with toluene and, finally, keep liners 15 min in methanol and rinse clean with methanol. Liners are dried, heated, and stored under inert gas and in sealed-glass tubes.
4. We describe the analysis of polar methanol and chloroform soluble metabolites without and in combination with the lipid metabolite complement. Major additional variants are selective enrichment of acidic or basic compounds, permutations of temperature, and extraction time for improved coverage of labile compounds and application of other solvents for selective extraction. Descriptions of alternate extraction protocols may be found elsewhere within this book.
5. The internal standard premixture for the analysis of polar compounds contains ribitol, 2,3,3,3-d_4-DL-alanine and D(-)-isoascorbic acid. Each component is prepared separately at 10 mg/mL in bidistilled water except for ribitol, which is dissolved in methanol. These stock solutions are combined into 50 mL bidistilled water and, thus, diluted to 0.02, 0.10, and 0.05 mg/mL final concentration, respectively. Diluted stocks can be stored at –20°C for a limited time. The internal standard solution for the analysis of the lipophilic metabolic complement needs to be freshly prepared and contains 2 mg/mL nonadecanoic acid methyl ester in chloroform. The internal standard premixtures can be extended to contain any set of stable isotope-labeled or synthetic internal standards.
6. The extraction mixture for plant material contains 300 parts methanol, 30 parts internal standard premixture for the polar metabolic complement (*see* **Note 5**), and 30 parts of the internal standard premixture for the lipophilic metabolic complement.
7. Back-up samples for in-line or manual derivatization can easily be generated by preparing additional aliquots from the surplus extracts and subsequent vacuum concentration. Note that rotors R96-13 and R120-111 require customized adaptors to accommodated tapered GC vials. Disposable 10-mL pipet tips, which are cut down to fit, may serve the same purpose.
8. The extraction mixture for yeast intercellular metabolites contains 350 parts methanol, 12 parts internal standard premixture for the polar metabolic complement (*see* **Note 5**), and 12 parts internal standard premixture for the lipophilic metabolic complement.
9. The reagent volumes of the in-line derivatization steps are adjusted to 10-µL sample volume. Increased sample volumes may not be fully redissolved in the 10-µL volume of methoxyamination reagent and result in nonmethoxyaminated but subsequently per-silylated side products, such as silylated hexopyranoses. The source of these side products is residual dried extract that sticks to the walls of the GC vials. These dried residues are not accessible through high-intensity shaking by the CTC agitator–incubator oven, but do not present a problem during manual agitation. For automated processing of extract, the aliquot volumes must not exceed 10 µL and need to be deposited at the bottom of the vial before vacuum centrifugation.
10. For continuous operation the GC–MS program needs to last less than 45 min. It is essential to either operate the GC–MS system under constant ambient temperature or check that increased ambient temperature owing to seasonal changes does not unexpectedly prolong the GC cycle time resulting in extended cooling times.
11. For the complete process of in-line derivatization a single syringe is used. This set up puts high demands on syringe cleanliness and mechanical performance. We mount a 25-µL syringe for best mechanical robustness of plunger and needle. Reagent and sample cross-contaminations may occur with inadequate wash protocols. Major contaminants from microbial and plant extracts are disaccharides, such as sucrose and trehalose or lipids and chlorophyll. When permanently present at high concentrations, these compounds are best removed by sequential treatment with polar and apolar solvents. The type of syringe cleaning cycle is best adjusted to the subsequent syringe task. We use hexane immediately before transferring MSTFA reagent and discard each first draw from the MSTFA reagent reservoirs taking care not to contaminate the reagents. Syringes are cleaned by maximum volume draws from the ethylacetate and *n*-hexane reservoirs.

Acknowledgments

The authors would like to thank Professor Lothar Wilmitzer and Dr. Oliver Fiehn, Max-Planck Institute of Molecular Plant Physiology (MPI-MP), Potsdam, Germany, and Professor Le Tran Binh, Institute of Biotechnology (IBT), Hanoi, Vietnam, for valuable advice, encouragement, and discussions. This work was supported by the Max-Planck society and the Bundesministerium für Bildung und Forschung (BMBF), grant PTJ-BIO/0312854.

References

1. Bino, R. J. Hall, R. D, Fiehn, O., et al. (2004) Potential of metabolomics as a functional genomics tool. *Trends Plant Sci.* **9,** 418–425.
2. Fernie, A. R., Trethewey, R. N., Krotzky, A. J., and Willmitzer, L. (2004) Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.* **5,** 763–769.
3. Wagner, C., Sefkow, M., and Kopka, J. (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOFMS metabolite profiles. *Phytochem.* **62,** 887–900.
4. Colebatch, G., Desbrosses, G., Ott, T., et al. (2004) Global changes in transcription orchestrate metabolic differentiation during symbiotic nitrogen fixation in *Lotus japonicus*. *Plant J.* **39,** 487–512.
5. Birkemeyer, C., Luedemann, A., Wagner, C., Erban, A., and Kopka, J. (2005) Metabolome analysis: the potential of *in vivo* labeling with stable isotopes for metabolite profiling. *Trends Biotechnol.* **23,** 28–33.
6. Fiehn, O., Kopka, J., Trethewey. R. N., and Willmitzer., L. (2000) Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal. Chem.* **72,** 3573–3580.
7. Roessner, U., Wagner, C., Kopka, J., Trethewey, R. N., and Willmitzer, L. (2000) Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J.* **23,** 131–142.
8. Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R. N., and Willmitzer, L. (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* **18,** 1157–1161.
9. Roessner, U., Luedemann, A., Brust, D., et al. (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* **13,** 11–29.
10. Kovàts. E. S. (1958) Gas-chromatographische Charakterisierung organischer Verbindungen: Teil 1. Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. *Helv. Chim. Acta* **41,** 1915–1932.
11. van Deursen, M. M., Beens, J., Janssen, H. -G., Leclercq, P. A., and Cramers, C. A. (2000) Evaluation of time-of-flight mass spectrometric detection for fast gas chromatography. *J. Chromatogr. A* **878,** 205–213.
12. Kopka, J., Schauer, N., Krueger, S., et al. (2005) GMD@CSB.DB: The Golm Metabolome Database. *Bioinformatics* **21,** 1635–1638.
13. Ausloos, P., Clifton, C. L., Lias, S. G., et al. (1999) The critical evaluation of a comprehensive mass spectral library. *J. Am. Soc. Mass Spectrom.* **10,** 287–299.
14. Stein, S. E. (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/ mass spectrometry data. *J. Am. Soc. Mass Spectrom.* **10,** 770–781.
15. Kopka, J., Fernie, A. R., Weckwerth, W., Gibon, Y., and Stitt, M. (2004) Metabolite profiling in plant biology: platforms and destinations. *Genome Biol.* **5,** 109–117.
16. Weckwerth, W., Wenzel, K., and Fiehn, O. (2004) Process for the integrated extraction, identification and quantification of metabolites, proteins and RNA to reveal their co-regulation in biochemical networks. *Proteomics* **4,** 78–83.
17. De Koning, W. and van Dam, K. (1992) A method for the determination of changes in glycolytic metabolites in yeast on a subsecond time scale using extraction at neutral pH. *Anal. Biochem.* **204,** 118–123.
18. Gonzalez, B., Francois, J., and Renaud, M. (1997) A rapid and reliable method for metabolite extraction in yeast using boiling buffered ethanol. *Yeast* **13,** 1347–1355.
19. Castrillo, J. I., Hayes, A., Mohammed, S., Gaskell, S. J., and Oliver, S. G. (2003) An optimized protocol for metabolome analysis in yeast using direct infusion electrospray mass spectrometry. *Phytochemistry* **62,** 929–937.
20. Batagelj, V. and Mrvar, A. (1998) Pajek: program for large network analysis. *Connections* **21,** 47–57.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Journal of Chromatography A, 993 (2003) 89–102
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

Comprehensive chemical derivatization for gas chromatography–mass spectrometry-based multi-targeted profiling of the major phytohormones

Claudia Birkemeyer, Ania Kolasa, Joachim Kopka*
Department Willmitzer, Max-Planck-Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm, Germany

Received 22 July 2002; received in revised form 10 February 2003; accepted 14 February 2003

Abstract

In the present investigation we report selection of the *N*-methyl-*N*-(*tert*.-butyldimethylsilyl)trifluoroacetamide (MTBSTFA) reagent as the most comprehensive derivatization protocol among 17 tested reactions covering trifluoroacetylation, pentafluorobenzylation, methylations, and trimethylsilylations. MTBSTFA allowed easy and robust *tert*.-butyldimethylsilyl derivatization of 1-aminocyclopropane-1-carboxylic acid, indole-3-acetic acid, (±)-jasmonic acid, salicylic acid, (±)-abscisic acid, *meta*-topolin, and *trans*-zeatin. Detection limits as analysed by selected ion monitoring quadrupole GC–MS were 0.2, 0.01, 1.0, 0.02, 0.3, 0.3, and 0.9 pmol of injected substance, respectively. Analysis of gibberellic acid A3, *trans*-zeatin riboside and (±)-abscisic acid-β-D-glucopyranosyl ester was best when coupled by splitting extracts and trimethysilylation. The MTBSTFA derivatization protocol was optimised, and validated. The preparation was insensitive to 2% residual water and to ≤1 day storage at room temperature. The final scheme was highly reproducible and successfully applied to extracts from ~300 mg (fresh mass) of tobacco (*Nicotiana tabacum*) root and *Arabidopsis thaliana* seedling.

Keywords: Derivatization, GC; *Nicotiana tabacum*; *Arabidopsis thaliana*; Tobacco; Organic acids; Plant hormones; Auxin; Cytotoxin

1. Introduction

Identification of auxin, the first phytohormone discovered by Went in 1928 [1], spurred a strong and lasting interest in fundamental research on plant growth regulators and applications in biotechnology. In succession abscisic acid, gibberellins, cytokinins, ethylene, jasmonic acids and salicylic acid were described, identified and demonstrated to exhibit respective regulatory functions. Even recently novel signalling substances such as brassinolides and the oligopeptide systemin [2] were found in plants.

Past and recent analysis of phytohormone action led to the emergence of the concept that none of the crucial biological functions, for example growth rate, growth orientation, development, and water balance, could be completely explained in a mono-causal manner. In contrast interplay of phytohormone levels nowadays appears to be more important to our understanding of phytohormone function than absolute concentrations of any single substance [3,4]. This novel insight was the incentive for our effort to

*Corresponding author. Tel.: +49-331-567-8262; fax: +49-331-5678-98262.
E-mail address: kopka@mpimp-golm.mpg.de (J. Kopka).

0021-9673/03/$ – see front matter
doi:10.1016/S0021-9673(03)00356-X

establish multi-targeted phytohormone profiling as an extension to our recently introduced technology of systems analysis, the GC–MS profiling of primary metabolites [5,6].

Today novel developments in quantitative phytohormone analysis are directed at either multi-parallel analysis or at increased sensitivity without compromising selectivity of detection. Downscaled sample requirement will increase spatial resolution phytohormone analysis. In contrast, multi-parallel analysis will allow novel insights into the interplay of phytohormone action. Our final goal is the efficient, sensitive, and comprehensive multi-targeted quantification of phytohormones from a single sample.

Several publications have already addressed the challenge of developing a suitable method for phytohormone profiling based on instrumental analytical technologies [7–14]. Phytohormones, like most constituents of signal transduction pathways, are trace compounds. Thus phytohormone analysis is subject to the common complications in trace analysis, namely laborious multi-step clean-up procedures, strong influence of sample matrix and ambient conditions [15]. The analytical platform of choice was gas chromatography coupled to mass spectrometry because of unsurpassed instrumental versatility, selectivity, sensitivity, and long-standing previous application in phytohormone analysis. Novel coupling technologies like solid-phase micro extraction, GC–GC coupling, and MS–MS techniques extend the already ample instrumental toolbox towards further means of micro-concentration and micro-separation.

Appropriate and stable derivatization of non-volatile compounds is crucial for successful GC analysis. Indeed nearly all major classes of phytohormones comprise polar compounds with high boiling points. A wide range of derivatizing protocols are available from comprehensive compendium guides [16,17]. Some have already been successfully applied to analysis of different phytohormone classes. Trifluoroacetylation was used in cytokinin analysis [18]. Trimethylsilylation was applied to cytokinin [19] and auxin [20] analysis. *tert.*-Butyldimethylsilylation of cytokinins was reported previously [21]. Alkylation with pentafluorbenzylbromide was successfully applied to the quantification of cytokinin [22] and auxin [23]. Methylation with diazomethane was reported in publications on jasmonic acid [24], auxin, salicylic acid, and abscisic acid [14]. Two-step procedures consisting of alkylation with diazomethane and subsequent trimethylsilylation were described for auxin [25] and gibbberellins [26]. A brief summary of further analytical methods developed for the quantification of the major phytohormones can be found in Ref. [27].

In the present study we reinvestigated and compared those chemical modification schemes which are in frequent use for the GC–MS analysis of phytohormones and which appeared to be versatile. In order of priority, the tested reagents were selected according to ease of handling, comprehensiveness of derivatization, and molar response ratio of the derivatives. The most promising scheme of a multi-parallel analysis was further optimised, validated, and standardised with a representative selection of phytohormones and other chemically related reference substances. Finally, we applied our method to plant matrices using a previously published extraction and clean-up procedure [14]. We introduce a sensitive, robust and easy-to-handle derivatization scheme appropriate for routine analysis of the major phytohormone classes from single plant samples.

2. Experimental

2.1. Standards and reagents

1-Aminocyclopropane-1-carboxylic acid (ACC; CAS 22059-21-8), myo-inositol (INO; CAS87-89-8), (±)-jasmonic acid (JA; CAS 3572-66-5), DL-tryptophan (Trp; CAS 54-12-6), gibberellic acid A3 (GA3; CAS 77-06-5), 5α-cholestane (CH; CAS 481-21-0), *n*-nonadecane (CAS 629-92-5), DL-α-tocopheryl acetate (CAS 7695-91-2) and the pesticide standard mixtures 8081 and EPA 508/508.1 were purchased from Sigma–Aldrich, Munich, Germany; *meta*-topolin (mT) and 24-epibrassinolide (BL; CAS 78821-43-9) were ordered from Duchefa, Haarlem, Netherlands; *trans*-zeatin (Z; CAS 1637-39-4), indole-3-acetic acid (IAA; CAS 87-51-4) and salicylic acid (SA; CAS 69-72-7) were from Merck, Darmstadt, Germany; (±)-abscisic acid (ABA; CAS 14375-45-2), (±)-abscisic acid-β-D-glucopyranosyl

ester (ABA-GE) and *trans*-zeatin riboside (ZR; CAS 6025-53-2) were received from Apex Organics Ltd., Honiton, UK. Where available, chemical abstracts system (CAS) registry numbers of the reference substances are provided.

The reagents were purchased as follows: *N*-methyl-*N*-(*tert*.-butyldimethylsilyl)trifluoroacetamide (MTBSTFA), *N*-methyl-*N*-(trimethylsilyl)heptafluorobutyramide (MSHFBA), *N*-methyl-*N*-(trimethylsilyl)trifluoroacetamide (MSTFA), *N*,*N*-dimethylformamidedimethylacetale (DMF-DMA), and trimethylsulphonium hydroxide (TMSH) were from Macherey-Nagel, Düren, Germany. *N*,*O*-Bis-(trimethylsilyl)acetamide (BSA), pentafluorobenzylbromide (PFBBr), *N*,*O*-bis(trimethylsilyl)trifluoroacetamide (BSTFA), *N*-trimethylsilyl-imidazole (TSIM), trimethylchlorosilane (TMCS), hexamethyldisilazane (HMDS), *N*-methylbis(trifluoroacetamide) (MBTFA), trimethylphenylammoniumhydroxide (TMAH), and methyl iodide (MeI)–potassium carbonate were ordered from Fluka, Buchs, Switzerland. Pyridine, methanol, dichlormethane (DCM), chloroform, ethylacetate (EtOAc) and formic acid, all HPLC grade, were supplied by J.T. Baker, Philipsburg, NJ, USA. Diazomethane was synthesised as described by Schlenk and Gellerman [28].

2.2. Sample preparation

2.2.1. Derivatization protocols

Standard stock solutions for the comparison of derivatization protocols were prepared in methanol at concentrations of 1 mg/ml. Only Z required addition of 1% (v/v) formic acid. A 5-μl sample of a 1:2 (v/v) dilution of each reference substance was combined with an equal volume of a 1:10 dilution of the 5α-cholestane stock solution, dried under nitrogen, and subjected to the derivatization procedures described below. Final amounts of 0.083 μg of each reference substance were used for GC–MS analysis. In the experiment addressing possible side product formation of MSTFA, BSTFA, MTBSTFA and MSHFBA reactions 0.166 μg SA was used (Table 2). Reaction parameters, e.g. solvent, volume ratios, incubation time and temperature, were not optimised for the initial screening. Instead general manufacturer's recommendations were applied unless indicated otherwise.

2.2.1.1. Trifluoroacetylation with MBTFA

Dissolve in 100 μl EtOAc, add 25 μl reagent, and heat to 120 °C for 2 min before analysis.

2.2.1.2. Methylation with diazomethane [28]

Add saturated ethereal solution of diazomethane, until yellow color is persistent, evaporate sample under nitrogen, and dissolve in 100 μl chloroform for GC–MS.

2.2.1.3. Methylation with DMF-DMA

Dissolve in 100 μl EtOAc, add 1000 μl DMF-DMA in pyridine 1:1 (v/v), and inject for GC–MS analysis when the solution becomes clear after 0.5–3 min.

2.2.1.4. Methylation with TMAH

Dissolve in 100 μl EtOAc, add 1 μl reagent, incubate for 10 min, and analyse directly by GC–MS.

2.2.1.5. Methylation with MeI–potassium carbonate [29]

Dissolve in 40 μl MeI–EtOAc (1:1, v/v), add 1–2 mg potassium carbonate, incubate for 1 h at 90 °C, and inject clear supernatant.

2.2.1.6. Methylation with TMSH

Dissolve in 100 μl EtOAc, add 50 μl reagent, incubate for 10 min at 100 °C and analyse by GC–MS.

2.2.1.7. Silylation with TSIM, MSTFA, BSTFA, MTBSTFA, and MSHFBA [30]

Add 100 μl reagent, incubate for 30 min at 90 °C, and analyse by GC–MS.

2.2.1.8. Silylation with HMDS–TMCS–pyridine (3:1:9, v/v/v) [31]

Dissolve in 100 μl of 3:1:9 (v/v/v) reagent mixture and incubate for 30 min before analysis by GC–MS.

2.2.1.9. Silylation with BSA–TSIM–TMCS, HMDS–TMCS (1:1, v/v) and HMDS–TMCS–pyridine (1:1:1, v/v/v) [31]

Dissolve in 100 μl of 1:1:1 (v/v/v) reagent mixture and incubate 1 h before analysis by GC–MS.

All procedures were carried out in at least three replications and performed at room temperature if not indicated differently.

2.2.2. Optimisation of the MTBSTFA protocol

Analysis of reproducibility, incubation time, incubation temperature, and the search for an internal standard substance with improved performance were carried out with 5 μl of 0.5-mg/ml stock solutions of each reference substance, *n*-nonadecane, and DL-α-tocopheryl acetate which were combined with 3 μl of a 0.1-mg/ml solution of 5α-cholestane. Samples were dried under a stream of nitrogen and incubated in 25 μl MTBSTFA reagent prior to quadrupole GC–electron ionization impact (EI)-MS analysis. Incubation was checked at 40, 60, 80, 100, and 120 °C temperature and at 30, 60, 120, and 180 min reaction time ($n=7$). All subsequent experiments were performed with optimised conditions, namely 1 h at 100 °C. Pesticide standard mixtures 8081 and EPA 508/508.1 were tested for potential candidate standard substances by adding 5 μl of each commercial preparation after heating.

Samples for storage stability tests were prepared as described above in 25 μl MTBSTFA and sealed in GC vials until further analysis. Storage was either at room temperature, 20–25 °C, or at −20 °C. A parallel set was sealed with 0.5 μl water added prior to incubation. Three replications of each set were analysed 5 h, 12 h, 24 h, 3 days, 7 days, and 14 days after start of incubation. For analysis of variance (ANOVA) the 5–24-h measurements were combined into the level "≤1 day" of the factor storage time, while the results of days 3–14 comprised the alternate level. Analysis of variance was performed with the statistical software package S-Plus 2000 standard edition release 3 (Insightful, Seattle, WA, USA).

Calibration curves and limits of detection were performed using multiple-component samples which were prepared by dilution of independent stock solutions.

2.2.3. Phytohormone profiling

Tobacco, *Nicotiana tabacum* cv. *Samsun*, plants were grown in sand under optimum growth chamber conditions. Roots were harvested 3 months after germination, rinsed under tap water until free of sand, and were then snap frozen in liquid nitrogen. *Arabidopsis thaliana* seedlings were germinated under sterile conditions on solid support and harvested after 2–3 weeks. For the purpose of this investigation representative batches were sampled, homogenised in a mortar under liquid nitrogen, and stored at −80 °C. Then 300 mg frozen fresh mass of these samples were extracted in 10 ml/g fresh mass of Bieleski solvent pre-cooled to −20 °C [32].

Co-purification of phytohormones from plant extracts was done exactly as described previously [14], except omitting the silica-based aminopropyl purification step. No further attempts at optimisation were undertaken. The final preparation was concentrated by vacuum centrifugation, 1 min at 200 mbar followed by 10 mbar to dryness. The optimised reaction protocol was applied and the MTBSTFA derivatives were analysed by quadrupole GC–EI-MS in the selected ion monitoring (SIM) mode and ion trap GC–chemical ionization (CI)-MS–MS in MS–MS reaction monitoring mode.

2.3. GC–MS analyses

GC–MS systems used in the present work were (i) an MD 800 GC–MS system (ThermoFinnigan, San Jose, CA, USA) with quadrupole technology supplied with split/splitless injection and MassLab software Version 1.4 and (ii) an ion trap Saturn 2000 GC–MS system (Varian, Palo Alto, CA, USA) supplied with programmed temperature vaporization injection and Saturn workstation software version 5.4. AMDIS software was employed to support peak finding before quantitative analysis and for automated deconvolution of reference mass spectra [33]. Identification of derivatives and side products was performed by co-chromatography and mass spectral fragmentation. The identification was supported by comparison to mass spectra presented in Ref. [34] as well as to a commercial mass spectral library in NIST98 format [35]. The quadrupole system was chosen for the analysis of side product formation (Table 2), because mass spectral deconvolution

software in combination with this technology allowed improved automated detection and better mass spectral comparisons with available libraries as compared to ion trap recordings. Information on a data file in the interchange format for AMDIS and NIST98 containing all mass spectra mentioned in Table 2 and respective ion trap mass spectra is to be found in Appendix A as supplementary data for cross-referencing.

Quadrupole GC–MS chromatograms were monitored by electron impact ionisation and either total ion monitoring, m/z 40–600, or in the experiments performed to determine calibration curves and detection limits via segmental selective ion monitoring (GC–EI-SIM-MS). Selected fragments were m/z 272 (ACC), m/z 232 (IAA), m/z 133 (JA), m/z 309 (SA), m/z 190 (ABA), m/z 469 (mT), m/z 302 (Z), and m/z 217 (5α-cholestane).

During initial analyses of derivatization protocols the ion trap mass spectrometer was operated in the EI-MS mode with total ion monitoring, m/z 40–600. Phytohormone profiles of plant samples were monitored in the CI-MS–MS mode with methanol reactant gas and positive ion detection. Maximum reaction time was 128 ms, maximum ionisation time 2 ms, scan rate 0.38 s/scan, multiplier offset 300 V, and emission current 30 μA, and the resonant waveform type was adjusted to MS–MS mode with a parent ion selection window of three atomic mass units. Parent ion selections and excitation amplitudes were segmental and changed as follows: ACC $[M+H]^+$, m/z 330, excitation amplitude 0.6 V; SA $[M+H]^+$, m/z 367, excitation amplitude 0.6 V; JA $[M+H]^+$, m/z 325, excitation amplitude 0.5 V; IAA $[M+H]^+$, m/z 290, excitation amplitude 0.5 V; ABA $[M-H_2O+H]^+$, m/z 361, excitation amplitude 0.6 V.

Arylene type 5% phenyl–95% methylpolysiloxane fused-silica capillary columns were chosen. A 30 m Rtx-5Sil MS fused-silica column, 0.25 mm inner diameter, 0.25 μm film thickness, supplied with a 10 m guard column (Restek, Bad Homburg, Germany) was used for the tests of different reagents and derivatization protocols. The GC–MS system was preconditioned each time the reagents were changed. Optimisation of the MTBSTFA protocol and phytohormone profiles of plant samples were performed without changes in performance on less expensive 30 m DB 5-MS fused-silica columns with 0.25 mm inner diameter, 0.25 μm film thickness (Agilent Technologies, Waldbronn, Germany).

Injection was hot splitless at 230 °C with an oven temperature ramp of 6 °C/min from 70 to 350 °C, ion source temperature was set to 230 °C, and transfer line was at 260 °C. Helium carrier gas was used at a flow-rate of 1 ml/min. These settings were used for all reagents and represent a compromise of previously described analyses [18–26]. The GC method was designed to cover a high temperature range and when tested still separated at least two derivatives from a commercial (±)-jasmonic acid isomer mixture.

The Saturn 2000 System was operated with a temperature program for controlled vaporization after injection, 0.5 min at 110 °C followed by a 250 °C/min ramp to 230 °C.

2.4. Definitions and calculations

Response was defined as chromatographic peak areas derived from mass spectrometric total ion, selected ion, or MS–MS recordings. Molar response was calculated as the quotient of analyte response over mole of substance injected into the GC–MS systems. Molar amount of injected substance was estimated by initial weight, dilution factor before derivatization, final volume of derivative and volume injected. Molar response ratios were the quotients of the molar responses of reference substances and a non-derivatized internal standard substance like 5α-cholestane. The MTBSTFA reaction procedure was optimised (Fig. 1) and tested for robustness by monitoring the relative yield of each derivative. The relative yield was calculated for each phytohormone as percentage of the maximum molar response ratio of the respective main derivative.

3. Results and discussion

3.1. Comparison of derivatization protocols

The reference substances for the following investigations were selected to cover most phytohormone classes by a single commercially available and affordable, naturally occurring compound. Thus the

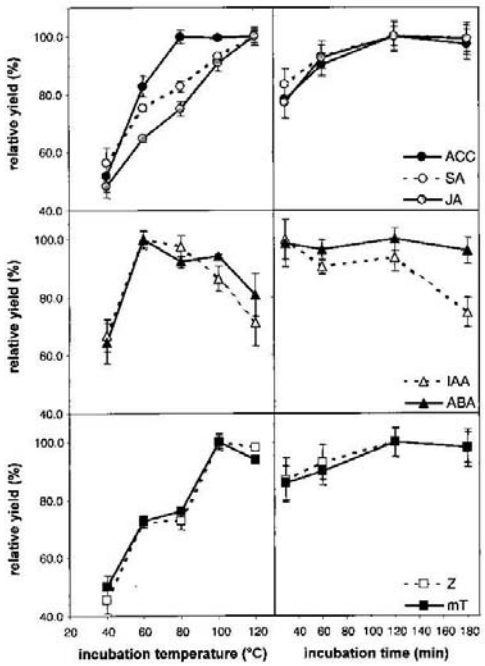


Fig. 1. Optimisation of the MTBSTFA reaction protocol. Relative yield was defined as % of maximum molar response ratio of each main derivative when monitored by quadrupole GC–EI-MS in total ion monitoring mode. Amounts per analysis were 2.5 μg of reference substance and 1.5 μg of 5α-cholestane. Permutation of incubation temperature ($n=7$); permutation of incubation time ($n=7$).

seven major phytohormone classes were represented by IAA, JA, SA, ABA, Z, mT, GA3, and BL (Table 1). Systemin had to be excluded because GC–MS technology is clearly unsuited for the analysis of oligo-peptides. The ethylene precursor, ACC, was included instead of the gaseous phytohormone. Furthermore we attempted to represent common phytohormone conjugates and typical functional groups by the two reference substances, ABA-GE and ZR. Reference substances of the equally important amino acid and inositol conjugates were not commercially available. Therefore, we included Trp and INO in order to assess their respective chemical behavior.

The comparative analyses of derivatization reactions necessitated a common substance for internal volume standardization. This substance was required to be inert with respect to all tested reagents. Therefore, initial experiments focused on the use of a range of hydrocarbons (data not shown). 5α-Cholestane was the best choice available with respect to inertness, intermediate volatility and distinctive fragmentation.

In Table 1 we summarise the results of the initial screening of reagents and protocols, among those listed in Section 2.2.1. In cases of multiple derivatives only those with highest molar response ratio are shown. IAA, JA, and ABA were easily detected by all protocols, but no strategy of chemical modification allowed analysis of all reference substances. Best coverage and molar response ratios were obtained with silylating reactions. Combined silylation and methylation increased the number of side products but allowed detection of BL. However, cytokinins, ABA-GE, and ZR were lost. This observation was also made for all stand-alone methylation reactions. Trifluoroacetylation exhibited strong selectivity and low molar response ratios. PFBBr was communicated as a highly sensitive reagent and robust to residual water [36]. In addition, pentafluorobenzylation is ideally suited for negative chemical ionisation (NCI)-MS [16,22]. In our hands PFBBr-derivatives exhibited high sensitivity even when monitored with EI-MS, but allowed detection only of ACC, IAA, JA, SA, ABA, and Trp.

Four silylating reagents with high donor strength, BSTFA, MSTFA, MSHFBA, and MTBSTFA, the later transferring *tert*.-butyldimethylsilyl groups, appeared to be most comprehensive. Comparison with less reactive silylating reagents under mild reaction conditions demonstrated that high reactivity was essential for this observation. Therefore, only those highly reactive reagents were further investigated with regard to the formation of side products and compared in a single large-scale experiment using a quadrupole GC–EI-MS system (Table 2). Most molar response ratios were increased as compared to previous ion trap results (Table 1). This effect was caused to a large extend by a reduced molar response of 5α-cholestane. Interestingly the quadrupole GC–MS system appeared also to discriminate the INO derivative and to be more sensitive to the derivatives of mT, Z, Trp, and ACC. The overall relative standard deviation (RSD) of these experiments was

Table 1
Molar response ratios of the main derivative obtained from 0.083 μg of each reference substance

Reagent[c]	Reference substance[a]												
	ACC	IAA	JA	SA	ABA	ABA-GE	mT	Z	ZR	GA3	BL	Trp	INO
A													
MTBSTFA	**0.146**	**0.237**	**0.149**	**0.270**	**0.262**		**1.042**	**0.440**	0.034	**0.169**		**0.376**	
B													
30–60 Min, room temperature:													
ESA–TMCS–TSIM (1:1:1, v/v/v)[b]		0.005											
HMDS–TMCS–pyridine (1:1:1, v/v/v)[b]		0.090	0.098	0.096	0.009					0.074			
HMDS–TMCS–pyridine (3:1:9, v/v/)[b]		0.073	**0.117**		0.007			0.021		**0.154**			
HMDS–TMCS (1:1, v/v)[b]	0.048	0.069	**0.115**	0.015	**0.121**			0.057		**0.588**			
30 Min, 90 °C:													
TSIM		0.005											
BSTFA	<0.001	**0.179**	**0.157**	**0.153**	**0.197**	0.010		0.044		**0.507**		0.009	**0.596**
MSTFA	0.069	**0.252**	0.096	0.098	**0.175**	0.020	0.014	**0.131**	**0.311**	**0.801**		**0.123**	**0.404**
MSHFBA	0.067	**0.231**	**0.133**	**0.149**	**0.146**	0.029	0.040	0.095	**0.343**	**0.691**		0.088	**0.231**
C													
MeI		0.066	0.026	<0.001						**0.138**			
Diazomethane		0.055	**0.133**	0.007	**0.146**					0.083	0.012		
DMF-DMA		0.050	0.088		**0.130**			0.002		0.088			
TMSH		**0.360**	0.056	0.001	**0.149**			0.066		0.009			
TMAH		**0.116**	**0.105**	0.002	**0.147**			0.009		0.072			
D													
Diazomethane–MSTFA	0.015	**0.179**	0.083	0.047	**0.139**			0.008		**0.643**	0.020		**0.184**
E													
METFA		0.087	0.051		0.004					0.012			
F													
FFBBr	**0.288**	0.086	**0.115**	**0.259**	**0.317**							**0.200**	

Molar response ratios were calculated from ion trap EI-MS total ion currents by normalisation to the signal of an equal amount of 5α-cholestane within each preparation. The table was compiled from multiple experiments. Each experiment was performed with aliquots of the same reference substance mixture ($n = 3$). Values >0.1 are in bold format. (A) *tert.*-Butyldimethylsilylation; (B) trimethylsilylations; (C) methylations; (D) combined trimethylsilylation and methylation; (E) trifluoroacetylation; (F) pentafluorobenzylation.

[a] ABA, (±)-Abscisic acid; ABA-GE, (±)-abscisic acid-β-D-glucopyranosyl ester; ACC, 1-aminocyclopropane-1-carboxylic acid; BL, 24-epibrassinolide; GA3, gibberellic acid A3; IAA, indole-3-acetic acid; INO, myo-inositol; JA, (±)-jasmonic acid; mT, *meta*-topolin; SA, salicylic acid; Trp, DL-tryptophan; Z, *trans*-zeatin; ZR, *trans*-zeatin riboside.

[b] Volume ratios.

[c] Refer to Section 2.1 for full identification of reagents.

Table 2
Molar response ratios of all observed derivatives generated from 0.083 μg of each reference substance

Reference substance[a]			TBS derivative			TMS derivative				
Compound	Molecular mass (g/mol)	Amount[c] (nmol)	Compound	No.	MTBSTFA	Compound	No.	BSTFA	MSTFA	MSHFBA
ACC	101.1	0.82	ACC TBS 1	2	0.688	ACC TMS 1	2	0.421	0.380	0.460
IAA	175.2	0.47	IAA TBS 1	2	0.005	IAA TMS 1	2	1.152	1.060	1.058
			IAA TBS 2	1	1.369	IAA TMS 2	1	0.140	0.002	0.001
JA	210.3	0.39	JA TBS 1	1	0.675	JA TMS 1	1	0.436	0.293	0.408
			JA TBS 2	2	0.002	JA TMS 2	2	0.002	0.163	0.010
SA	138.1	1.20[b]	SA TBS 1	2	1.464	SA TMS 1	2	0.613	0.473	0.699
ABA	264.3	0.31	ABA TBS 1	1	1.229	ABA TMS 1	1	0.834	0.746	0.693
			ABA TBS 2	2	0.050	ABA TMS 2	2	0.001	0.087	0.018
ABA-GE	426.5	0.19		–	–	ABA-GE TMS 1	(4)	0.023	0.008	0.015
mT	241.4	0.34	mT TBS 1	2	1.421	mT TMS 1	2	0.595	0.620	0.619
				–	–	mT TMS 2	1	0.030	0.039	0.025
				–	–	mT TMS 3	3	0.005	0.017	0.018
Z	219.2	0.38	Z TBS 1	2	1.856	Z TMS 1	2	1.058	1.057	0.966
			Z TBS 2	3	0.007	Z TMS 2	3	0.050	0.079	0.077
ZR	351.4	0.24	ZR TBS 1	(3)	0.016	ZR TMS 1	4	1.548	1.494	1.561
GA3	346.4	0.24	GA3 TBS 1	1	0.100	GA3 TMS 1	3	0.999	1.066	0.950
				–	–	GA3 TMS 2	3	0.081	0.077	0.057
BL	480.8	0.17		–	–		–	–	–	–
Trp	204.2	0.41	Trp TBS 1	3	0.005	Trp TMS 1	3	0.329	1.175	1.270
			Trp TBS 2	2	1.254	Trp TMS 2	2	0.053	0.191	0.083
				–	–	Trp TMS 3	1	0.030	<0.001	<0.001
				–	–	Trp TMS 4	2	0.160	<0.001	<0.001
INO	180.2	0.46		–	–	INO TMS 1	6	0.009	0.137	0.068

Molar response ratios were calculated from quadrupole EI-MS total ion currents of a single experiment ($n=3$) by normalisation to the signal of an equal amount of 5α-cholestane within each preparation. Mass spectra of all *tert.*-butyldimethylsilyl (TBS) and trimethylsilyl (TMS) derivatives mentioned in the table are available on request from the communicating author from the mass spectral library included as supplementary data in Appendix A. The number of trimethylsilyl groups (No.) is listed, brackets indicate an estimated number.

[a] ABA, (±)-Abscisic acid; ABA-GE, (±)-abscisic acid-β-D-glucopyranosyl ester; ACC, 1-aminocyclopropane-1-carboxylic acid; BL, 24-epibrassinolide; GA3, gibberellic acid A3; IAA, indole-3-acetic acid; INO, myo-inositol; JA, (±)-jasmonic acid; mT, *meta*-topolin; SA, salicylic acid; Trp, DL-tryptophan; Z, *trans*-zeatin; ZR, *trans*-zeatin riboside.

[b] 0.166 μg per preparation.

[c] Amount per analysis.

21, 31, 27, and 24% ($n=3$) including all minor products of the reactions with, MTBSTFA, BSTFA, MSTFA, and MSHFBA, respectively. The most comprehensive derivatization was trimethylsilylation. All trimethylsilyl reagents, BSTFA, MSTFA, and MSHFBA, generated a single main product and identical side products except for ABA-GE and BL. In the case of ABA-GE low signal intensity was likely caused by instability of the conjugate as judged by occurrence of free silylated glucose (data not shown). In comparison to trimethylsilyl reagents, MTBSTFA was slightly more sensitive and less prone to formation of side products (Table 2; refer to Trp and mT). In some cases MTBSTFA exhibited a preference for a lower degree of substitution, namely IAA and Trp. Unfortunately this property of

MTBSTFA did not allow analysis of GA3, ZR, ABA-GE, BL, or INO. In the case of GA3 and ZR we detected minor signals of derivatives with low degree of substitution, but the bulk derivative was lost.

In view of the ultimate goal of our efforts—the sensitive close to comprehensive multi-targeted quantification of phytohormones—we decided on an in-depth analysis of the MTBSTFA derivatization reaction. This decision took into account first the prospective sensitivity, namely the combined aspects of low side product formation, high molar response and low complexity of fragmentation. Secondly we expected higher selectivity of detection, because the mass spectral fragmentation pattern of MTBSTFA derivatives generates typical $[M-57]^+$ and $[M-15]^+$ fragments from, in most cases, still detectable molecular ions. Finally we took into account the purity of the reagent and the stability of derivatives [37]. The drawback of MTBSTFA derivatization, however, is low sensitivity for gibberellins and monosaccharide conjugates. This deficiency may be circumvented by splitting of extracts and in parallel alternative analysis by trimethylsilylation. Trimethylsilylation has a more comprehensive potential [5], but showed interference with residual water. Moreover, mass spectral fragmentation patterns were clearly more complex and less specific. Therefore a higher demand on pre-purification and concentration from plant matrices and thus lower overall sensitivity was expected.

3.2. Optimisation of the MTBSTFA protocol

3.2.1. Repeatability of GC–MS analysis

The MTBSTFA protocol was optimised using a quadrupole GC–EI-MS system because of the higher sensitivity in the EI-MS mode and comparative ease of handling and data processing. For this purpose we performed experiments of nine repeated injections in the course of 10 h with a reference mixture of ACC, IAA, JA, SA, ABA, mT, and Z and varying internal standard substances. The molar response ratios of the derivatives using our initial choice of the 5α-cholestane standard had 6.0–13.1% RSD (Table 3). In our hands this level of repeatability is typically achieved by GC–EI-MS systems, when MTBSTFA or MSTFA are used as solvents for injection.

In an attempt to assess possible improvement of GC–MS reproducibility we tested internal standardization by *n*-nonadecane, DL-α-tocopheryl acetate, and each of the components of the pesticide standard mixtures 8081 and EPA 508/508.1. The pesticide mixtures allowed the fast screening of a large range of different compound classes. which were in part derivatized by MTBSTFA. None of the tested compounds exhibiting higher as well as lower boiling points qualified for a better internal standard

Table 3
Repeatability of the molar response ratio of the main derivative synthesised from 0.083 μg of each reference substance

Reference substance[a]			Total ion response, molar response ratio	
Compound	Amount[b] (nmol)	Derivative	Average	RSD (%)
ACC	0.82	ACC TBS 1	1.004	6.0
IAA	0.47	IAA TBS 2	1.281	13.1
JA	0.39	JA TBS 1	0.857	6.2
SA	0.60	SA TBS 1	0.920	6.8
ABA	0.31	ABA TBS 1	1.438	8.7
mT	0.34	mT TBS 1	1.621	11.8
Z	0.38	Z TBS 1	2.040	13.0
Trp	0.41	Trp TBS 2	1.300	10.7
CH	0.48	–	–	–

Molar response ratios were calculated from quadrupole EI-MS total ion currents ($n=9$) set to a scanning range of m/z 40–600 using the signal of 5α-cholestane within each preparation for normalisation.

[a] ABA, (±)-Abscisic acid; ACC, 1-aminocyclopropane-1-carboxylic acid; CH, 5α-cholestane; IAA, indole-3-acetic acid; JA, (±)-jasmonic acid; mT, *meta*-topolin; SA, salicylic acid; Trp, DL-tryptophan; Z, *trans*-zeatin.

[b] Amount per analysis.

of any of the reference derivatives than 5α-cholestane. Therefore, we continued internal standardization with this compound.

3.2.2. Optimisation of reaction conditions

Fig. 1 summarises the effects of permuted incubation time and temperature on relative yield. The relative yield was calculated separately for the main derivatives of each phytohormone. The maximum molar response ratio obtained in each experiment was set to 100% relative yield. Side products did not accumulate under any of the conditions tested. Derivatives were grouped according to similarity of behavior. These groups were related to the GC elution sequence of derivatives. Cytokinins exhibited almost identical behavior, and IAA and ABA were highly similar, whereas JA, SA, and ACC showed a similar tendency. In general incubation temperature exhibited a stronger influence on reaction yield than incubation time. The optimum compromise for all phytohormones was 1-h incubation at 100 °C.

3.2.3. Stability and storage

An experiment in factorial 2^3 design was performed to detect parameters which might influence robustness of analysis after final derivatization. Two levels of three factors were investigated by 24 experiments comprising three replications of each possible combination of factors. The stability of the derivative was tested for a typical time of analysis, ≤1 day, as compared to storage for 3–14 days. Influence of residual water, a common problem in phytohormone preparations, was checked by addition of 2% (v/v) of water before reaction with MTBSTFA. Storage temperature was the third parameter tested. Typically samples are exposed to room temperature before GC analysis. Therefore, we compared storage at −20 °C with exposure to room temperature.

A three-way ANOVA was performed for each phytohormone. IAA, SA, and ABA analysis was not influenced by any of the tested challenges to robustness. In the case of IAA and ABA this observation was contrary to our expectations (Fig. 1). Each derivative was stable in the presence of trace amounts of water. All relative yields were 90–100%. Storage time increased relative yield of ACC by a factor of 1.5 ($P<0.001$) and relative yield of JA by a factor of 2.0 ($P=0.004$). In contrast, relative yield of mT was reduced by a factor of 0.65 ($P<0.001$). This finding may be indicative of incomplete derivatization in the case of ACC and JA and shows slight long-term instability of the mT derivative. Lowering storage temperature to −20 °C appeared not to be beneficial. In contrast the relative yield of Z was reduced by a factor of 0.9 ($P=0.006$). This effect was increased in combination with long storage time ($P<0.001$). No other interaction of factors was

Table 4
Detection limits of phytohormones expressed as amount required before derivatization by MTBSTFA[a]

Phytohormone	Selected ion monitoring					
	Derivative	Fragment (*m/z*)	Relative abundance (%)[c]	Detection limit[b] (ng)	Detection limit[b] (pmol)	*S/N*
ACC	ACC TBS 1	272	7.3	0.50	5.0	4:1
IAA	IAA TBS 2	232	19.4	0.05	0.3	5:1
JA	JA TBS 1	133	7.0	5.00	24.0	5:1
SA	SA TBS 1	309	18.5	0.05	0.4	15:1
ABA	ABA TBS 1	190	6.0	2.00	7.5	4:1
mT	mT TBS 1	469	8.9	2.00	8.3	5:1
Z	Z TBS 1	302	7.9	5.00	23.0	4:1

Quadrupole GC–EI-MS was set to selected ion monitoring. The signal-to-noise ratio at the limit of detection was calculated based on the maximum amplitude of the background signal in the vicinity of the respective peak.

[a] ABA, (±)-Abscisic acid; ACC, 1-aminocyclopropane-1-carboxylic acid; IAA, indole-3-acetic acid; JA, (±)-jasmonic acid; mT, *meta*-topolin; SA, salicylic acid; Z, *trans*-zeatin.

[b] Total derivatization volume was 25 μl. A 1-μl sample was applied to GC–MS analysis.

[c] The relative abundance of selected ion fragments was determined in parallel by total ion monitoring experiments with *m/z* 40–600.

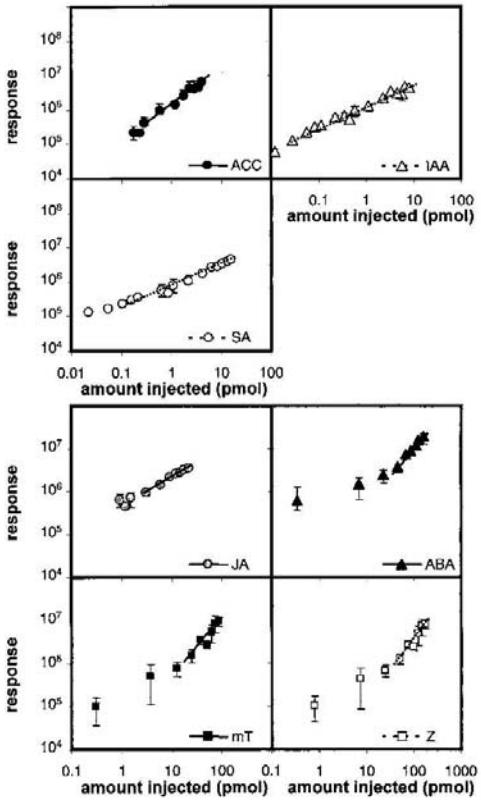


Fig. 2. Calibration curves ($n=6$) demonstrating the working range of the MTBSTFA derivatization protocol of 1-h incubation at 100 °C in a volume of 25 μl, as determined by quadrupole GC–EI-MS set to selected ion monitoring mode. The smallest amounts shown represent the detection limits. Fragments and the signal-to-noise ratios at the detection limits were as listed in Table 4. Error bars represent standard deviation.

detectable. ANOVA clearly demonstrated general robustness of the selected protocol. Furthermore, we checked effects on side product formation and found their relative occurrence to be invariant.

3.2.4. Calibration and limits of detection

Quadrupole GC–EI-MS set to selected ion monitoring mode was used for analysis of detection limits and respective signal-to-noise ratios (S/N). Only 1/25 of final the sample volume was analyzed by GC–MS operated with splitless injection. The noise value was determined as the maximum amplitude of the background signal in a range of ±5 times the respective peak width. Peak height was calculated from the average noise level to peak apex. The detection limits are presented as amount required before derivatization with MTBSTFA reagent (Table 4). Sensitivity varied within two orders of magnitude among the different compounds analysed. Detection of IAA and SA was highly sensitive, 0.3 and 0.4 pmol, respectively, whereas JA and Z exhibited detection limits of 24 and 23 pmol per sample. ACC, ABA, and mT had intermediate detection limits of 5.0, 7.5, and 8.3 pmol. The fragments chosen for selected ion monitoring had relative abundances of 7–19% and were mostly in the high-molecular mass range. Both high relative abundance and high mass of available fragments contributed to the considerable sensitivity of *tert.*-butyldimethylsilyl derivatives as compared to TMS derivatives.

Calibration curves of the phytohormones were determined and are presented in double logarithmic scale (Fig. 2). The smallest amount shown corresponds to the respective detection limits (Table 4). In the sub pmol to pmol range the reference substances, ACC, IAA, SA, and JA, all exhibited clear linear behavior. The response functions of ABA, mT, and Z were sigmoidal. Calibration curves which were extended into the nmol range all had a sigmoidal shape and indicated upper detection limits of 10–100 nmol injected.

3.3. Phytohormone profiling

The MTBSTFA protocol was successfully applied to analysis of extracts from tobacco root and seedlings of *A. thaliana*. A published extraction and purification method [14] was adopted and phytohormone fractions prepared accordingly from representative samples of ~0.3 g fresh mass. Analysis with quadrupole GC–EI-MS in selective ion monitoring mode exhibited in part intense peaks which were identified by spiking experiments. Moreover, each phytohormone was monitored using four different fragments in four consecutive runs. These experiments and total ion scanning analysis indicated inadequate sample purity for routine analysis with quadrupole GC–EI-MS. For the unequivocal demon-

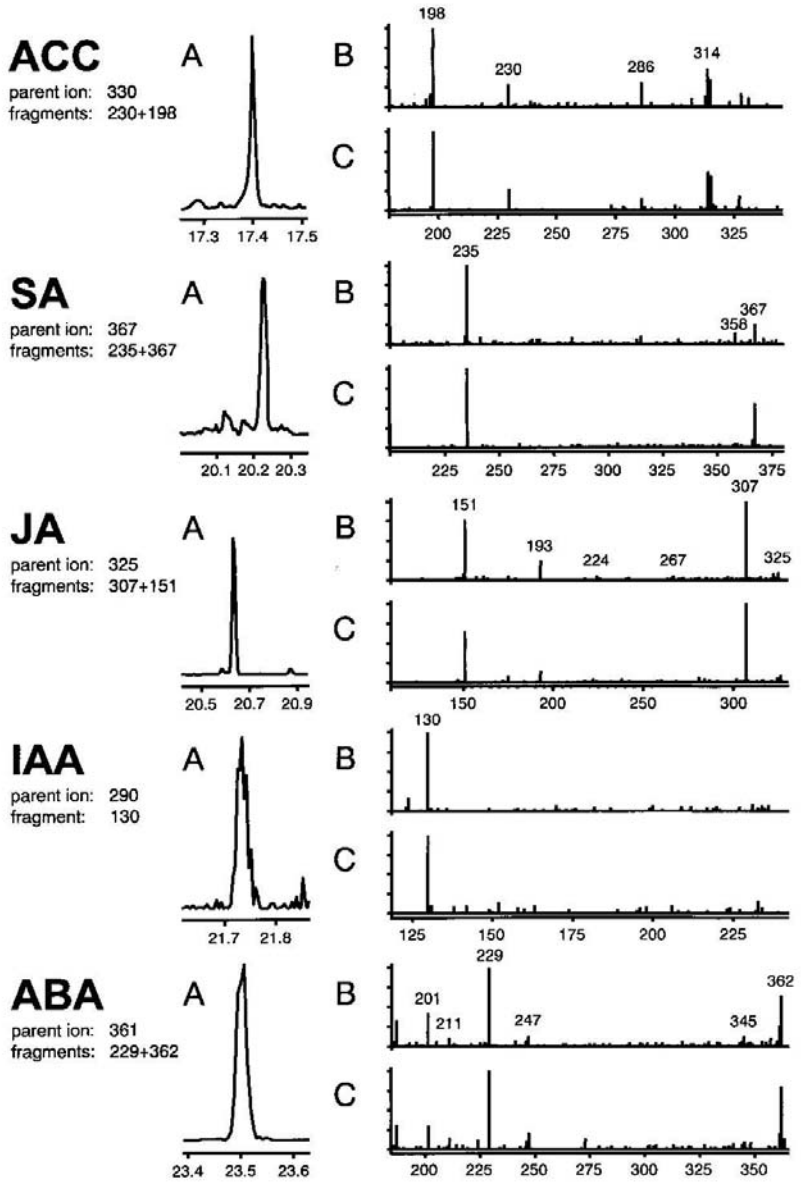


Fig. 3. MTBSTFA phytohormone profile of 0.3 g tobacco (*Nicotiana tabacum*) root, recorded with an ion trap GC system in CI-MS–MS mode. (A) Specific MS–MS fragment traces. MS–MS fragmentation of reference substances (B) are compared to the MS–MS spectra of endogenous plant compounds (C). MS–MS spectra were taken from the peak apexes. ABA, (±)-abscisic acid; ACC, 1-aminocyclopropane-1-carboxylic acid; IAA, indole-3-acetic acid; JA, (±)-jasmonic acid; SA, salicylic acid.

stration of the presence of endogenous phytohormones we employed ion trap GC–CI-MS–MS. Analysis of MS–MS spectra allowed identification of ACC, IAA, SA, JA, and ABA from both tobacco root (Fig. 3) and *A. thaliana* seedlings (data not shown). In shoot organs of *A. thaliana* IAA, SA, and JA, and ABA are typically found in concentrations of 100–1000 pmol/g fresh mass, whereas SA and ABA mostly range from 10 to 100 pmol/g fresh mass [14]. Our successful identification of IAA, SA, JA, and ABA was therefore in agreement with expectations. ACC was previously not noticed in similar preparations (Fig. 3). We did not try to monitor Z in this experiment because the Z concentration reported to occur in tobacco seedlings did not exceed 0.25 pmol/g fresh weight [38].

4. Conclusions

We present a novel method appropriate for comprehensive chemical derivatization and subsequent gas chromatography–mass spectrometry of phytohormones. The coverage of phytohormone classes is broader than reported previously for a single analysis [14]. However, current means of joined extraction and preparation of phytohormone fractions from plant samples restrict the potential of our analysis to five endogenous target compounds, e.g. ACC, IAA, SA, JA, and ABA. Our future efforts will focus on extending the preparation protocol in order to fully exploit our novel method. Currently ion trap GC–CI-MS–MS analysis is mandatory and quantitative standardisation is restricted to the use of stable isotope labelled reference substances.

Acknowledgements

The authors thank all members of the Max-Planck-Institute of Molecular Plant Physiology, Golm, Germany, especially Cornelia Wagner for patient support in performing GC–MS measurements, and Dr Alisdair Fernie and Professor Lothar Willmitzer for fruitful discussions and close scrutiny of our work. Our special gratitude belongs to Dr Hilde Boiten, Dr Els Prinsen (University of Antwerp, Belgium), Dr Axel Müller, and Professor Elmar W. Weiler (University of Bochum, Germany) for their readiness to give insight into their methods of phytohormone analysis and further invaluable support.

Appendix A. Supplementary file 1

The data file, phytoh.msp[1], contains mass spectra of all phytohormone derivatives mentioned in Tables 2–4. Quadrupole and ion trap electron impact ionisation mass spectra are included.

The spectrum name was designed to allow sorting according to the reference substance. For example, the name ABA TBS1≠EI≠Q≠MTBSTFA[2] codes for the name of the reference substance and type of derivative, mode of ionisation, mode of mass spectral detection and reagent. The spectrum ID allows sorting according to reagent, mode of ionisation, mode of mass spectral detection, and source chromatogram, for example MTBSTFA≠EI≠Q≠1235DW21.

References

[1] F.W. Went, Rec. Trav. Bot. Neerl. 25 (1928) 1.
[2] G. Pearce, D. Strydom, S. Johnson, C.A. Ryan, Science 253 (1991) 895.
[3] J.D.B. Weyers, N.W. Paterson, New Phytol. 152 (2001) 375.
[4] P.J. Davies (Ed.), Plant Hormones, Kluwer, Dordrecht, Netherlands, 1995.
[5] O. Fiehn, J. Kopka, P. Doermann, T. Altmann, R.N. Trethewey, L. Willmitzer, Nat. Biotechnol. 18 (2000) 1157.
[6] U. Roessner, C. Wagner, J. Kopka, R.N. Trethewey, L. Willmitzer, Plant J. 23 (2000) 131.
[7] E. Prinsen, W. van Dongen, E.L. Esmans, H.A. van Onckelen, J. Chromatogr. A 826 (1998) 25.
[8] M. Kowalczyk, G. Sandberg, Plant Physiol. 127 (2001) 1845.
[9] P.I. Dobrev, M. Kaminek, J. Chromatogr. A 950 (2002) 21.

[1] The file format *.msp may be imported into either NIST98 or AMDIS software (to be downloaded from http://chemdata.nist.gov/mass-spc/Srch_v1.7/index.html and http://chemdata.nist.gov/mass-spc/amdis/, respectively). The file phytoh.msp is available on request from the communicating author.

[2] EI, electron impact; MSTFA, *N*-methyl-*N*-(trimethylsilyl)trifluoroacetamide; MTBSTFA, *N*-methyl-*N*-(*tert*.-butyldimethylsilyl)trifluoroacetamide reagent; Q, quadrupole technology; T, ion trap technology.

[10] B.-F. Liu, X.-H. Zhong, Y.-T. Lu, J. Chromatogr. A 945 (2002) 257.
[11] E. Prinsen, P. Redig, M. Strnad, I. Galis, W. van Dongen, H. van Onckelen, Methods Mol. Biol. 44 (1989) 245.
[12] G. Schneider, J. Schmidt, J. Chromatogr. A 728 (1996) 371.
[13] W. Rademacher, J.E. Graebe, Ber. Deutsch. Bot. Ges. Bd. 97 (1984) 75.
[14] A. Mueller, P. Duechting, E.W. Weiler, Planta 216 (2002) 44.
[15] L. Rivier, A. Crozier (Eds.), Principles and Practice of Plant Hormone Analysis, Academic Press, London, 1987.
[16] D.R. Knapp, Handbook of Analytical Derivatization Reactions, Wiley, New York, 1979.
[17] K. Blau, J.M. Halket (Eds.), Handbook of Derivatives for Chromatography, Wiley, Chichester, 1993.
[18] M. Ludewig, K. Doerffling, W.A. Koenig, J. Chromatogr. 243 (1982) 851.
[19] L.M.S. Palni, R.E. Summons, D.S. Letham, Plant Physiol. 72 (1983) 858.
[20] J. Badenoch-Jones, R.E. Summons, B.G. Rolfe, D.S. Letham, J. Plant Growth Regul. 3 (1984) 23.
[21] C.H. Hocart, O.C. Wong, D.S. Letham, S.A.B. Tay, J.K. MacLeod, Anal. Biochem. 153 (1986) 85.
[22] D.S. Letham, S. Singh, O.C. Wong, J. Plant Growth Regul. 10 (1991) 107.
[23] E. Prinsen, S. van Laer, S. Oeden, H. van Onckelen, Methods Mol. Biol. 141 (2000) 49.
[24] O. Miersch, H. Bohlmann, C. Wasternack, Phytochemistry 50 (1999) 517.
[25] A. Edlund, S. Ekloef, B. Sundberg, T. Moritz, G. Sandberg, Plant Physiol. 108 (1995) 1043.
[26] S.J. Croker, P. Gaskin, P. Hedden, J. MacMillan, K.A.G. MacNail, Phytochem. Anal. 5 (1994) 74.
[27] G. Sembdner, G. Schneider, K. Schreiber, Methoden zur Pflanzenhormonanalyse, Springer, Berlin, 1987.
[28] H. Schlenk, J.L. Gellerman, Anal. Chem. 32 (1960) 1412.
[29] D.R. Knapp, in: Handbook of Analytical Derivatization Reactions, Wiley, New York, 1979, p. 155.
[30] J.L. Little, J. Chromatogr. A 844 (1999) 1.
[31] D.R. Knapp, in: Handbook of Analytical Derivatization Reactions, Wiley, New York, 1979, p. 566.
[32] R.L. Bieleski, Anal. Biochem. 9 (1964) 431.
[33] S.E. Stein, J. Am. Soc. Mass Spectrom. 10 (1999) 770.
[34] P. Gaskin, J. MacMillan, GC–MS of Gibberellins and Related Compounds: Methodology and a Library of Spectra, Cantocks Enterprises, Bristol, UK, 1991.
[35] P. Ausloos, C.L. Clifton, S.G. Lias, A.I. Mikaya, S.E. Stein, D.V. Tchekhovskoi, O.D. Sparkman, V. Zaikin, D. Zhu, J. Am. Soc. Mass Spectrom. 10 (1999) 287.
[36] H. Boiten, Department of Biology, University of Antwerp, personal communication, 2001.
[37] E.J. Corey, A. Venkateswarlu, J. Am. Chem. Soc. 94 (1972) 6190.
[38] T. Werner, V. Motyka, M. Strnad, T. Schmülling, Proc. Natl. Acad. Sci. USA 98 (2002) 10487.

Opinion TRENDS in Biotechnology Vol.23 No.1 January 2005

Metabolome analysis: the potential of *in vivo* labeling with stable isotopes for metabolite profiling

Claudia Birkemeyer, Alexander Luedemann, Cornelia Wagner, Alexander Erban and Joachim Kopka

Department Willmitzer, Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, D-14467 Golm, Germany

Metabolome analysis technologies are still in early development because, unlike genome, transcriptome and proteome analyses, metabolome analysis has to deal with a highly diverse range of biomolecules. Combinations of different analytical platforms are therefore required for comprehensive metabolomic studies. Each of these platforms covers only part of the metabolome. To establish multiparallel technologies, thorough standardization of each measured metabolite is required. Standardization is best achieved by addition of a specific stable isotope-labeled compound, a mass isotopomer, for each metabolite. This suggestion, at first glance, seems unrealistic because of cost and time constraints. A possible solution to this problem is discussed in this article. Saturation *in vivo* labeling with stable isotopes enables the biosynthesis of differentially mass-labeled metabolite mixtures, which can be exploited for highly standardized metabolite profiling by mass isotopomer ratios.

Introduction

The field of analytical biochemistry has recently received a novel extension: metabolomics, a concept that has been defined as the science of the comprehensive monitoring of the metabolic complement in biological systems [1–6]. Metabolomics provides information about biological systems which cannot be obtained by the classical '–omics' approaches: genomics, transcriptomics and proteomics. Metabolomics could be viewed, therefore, as the missing fourth Rosetta stone [7], which fills the metabolic gap within the previously developed systems-wide approaches towards the global analysis of biological processes.

Metabolomic approaches are under dynamic development and several synonyms have been suggested, such as metabonomics [8], metabolite profiling [9] or fingerprinting [4,6]. A multitude of analytical platforms has been introduced [10], including spectroscopy fingerprints at infrared (IR), near infrared (NIR) or ultraviolet (UV) wavelength ranges, gas chromatography–mass spectrometry (GC-MS) [2,3,9], liquid chromatography–electrospray ionization–mass spectrometry (LC-ESI-MS) [11–13], capillary electrophoresis coupled to mass spectrometry (CE-MS) [14] or liquid chromatography with nuclear magnetic resonance spectroscopy (LC-NMR) [8,15], just to mention a few. There is no single analytical platform currently conceivable that would enable the multiparallel analysis of the complete metabolome [10,16], which comprises the full range of chemically diverse biomolecules

Corresponding author: Kopka, J. (kopka@mpimp-golm.mpg.de).

Glossary

Dynamic range: The range of concentrations, between detection limit and maximum amount of a substance to be quantified by one analytical technology.

Hyphenated technologies: Hyphenation stands for the combination of at least two principles of chemical separation in a single instrument, such as gas chromatography–mass spectrometry (GC-MS), capillary electrophoresis–electrospray mass spectrometry (CE-ESI-MS) or high-performance liquid chromatography–nuclear magnetic resonance spectroscopy (HPLC-NMR). Exploiting two chemical properties for compound separation is prerequisite for in-parallel analysis of multiple compounds from a complex mixture.

Mass isotopomer: Chemical substances are composed of naturally occurring (or technically enriched) mixtures of atomic isotopes, which have, as a rule, the same chemical properties but exhibit different mass. A modern mass spectrometer can resolve the mass differences of a single isotope substitution within a molecule. Each mass variant of a chemical substance is called a mass isotopomer.

Mass isotopomer distribution: The mass isotopomer distribution of molecules can be precisely calculated and is dependent: (i) on the number of atoms present in a molecule and (ii) on the natural or technically enriched isotope abundances of each element [46]. High isotope enrichment and low atom numbers result in highly abundant, fully labeled mass isotopomers. Low enrichment and high atom numbers favor partial labeling and cause a broad distribution. For example, a four-carbon molecule might carry a ^{12}C- or ^{13}C-label at either of the positions. The chances for a fully ^{13}C-labeled four-carbon mass isotopomer at ambient 1.1% enrichment are negligible, $0.011^4 = 1.46 \times 10^{-8}$, whereas chances are intermediate, $0.70^4 = 0.24$, or high, $0.99^4 = 0.96$, at 70% and 99% isotope enrichment, respectively. By comparison, the chances of a fully labeled 30-carbon mass isotopomer at 99% enrichment are clearly reduced, $0.99^{30} = 0.74$.

Matrix effect: The matrix effect is a long-standing observation in chemical and enzymatic analyses of complex biological samples. Namely, the nature or composition of complex samples can influence the apparent amount of metabolites and thus might lead to false quantitative results. Matrix effects can either stabilize labile compounds (matrix stabilization) or suppress compound measurements (matrix suppression). Matrix effects can occur at any step during chemical analysis, from extraction through clean-up, to final instrumental analysis. Well known examples are the matrix suppression effects of electrospray ionization–mass spectrometry [28] or matrix-assisted laser desorption–time of flight–mass spectrometry [29] technologies or the oxidation of labile metabolites, such as vitamin C.

Recovery: Recovery measurements are the analytical means to control and standardize metabolite measurements for matrix effects. Recovery is routinely expressed as a percentage or ratio. The comparison made is between equal amounts of metabolites either supplied as a pure reference sample or added to the biological sample under scrutiny. Recovery analyses are most elegantly performed by using chemically synthesized mass isotopomers, which can be distinguished from the respective naturally occurring counterparts by high-resolution mass analysis.

Metabolic phenotype: The qualitative and quantitative inventory of all metabolites in a biological sample [2–4].

www.sciencedirect.com 0167-7799/$ - see front matter doi:10.1016/j.tibtech.2004.12.001

from low molecular weight volatiles to storage polymers, such as starch or triacylglycerols. The diversity of required methods is in stark contrast to genome, transcriptome and proteome profiling technologies, which monitor molecules of highly similar chemical properties, such as DNA, RNA and proteins, respectively.

Metabolome analyses not only need to accommodate the high diversity of biomolecules but also need to cover the vast dynamic range (see Glossary) of metabolite concentrations. These encompass highly abundant nutrients or primary metabolites and equally important trace compounds that might carry biological signals. In addition, the metabolome is formed by a complex network of reactions, which are subject to rapid enzymatic turnover [17–21]. Extreme care and fast inactivation of all biochemical reactions during sampling is therefore vital [2,12,22,23], which is not the case in proteome and transcriptome analyses. Although the sequence information embedded within protein and RNA structure enables unequivocal identification of the source organism, metabolites *per se* do not carry information on their respective origins. Thus, metabolite measurements need to be controlled for artifact chemical contamination that might arise during biological experimentation or chemical analysis.

Even though there appear to be considerable technical obstacles, metabolome analyses are in high demand and have been widely proposed for studies in molecular physiology [2,16,24], functional genomics [1,3,7,15], clinical chemistry [8], biomarker discovery, research on the mode of drug action and monitoring of drug therapy [15,25,26]. This interest necessitates a short discussion of the properties of novel metabolomic approaches compared with classical chemical analytics.

Analytical approaches of metabolome analysis: general variants, properties and applications

Four major variants of analytical approaches are currently conceivable, fingerprinting, profiling, absolute quantification of pool sizes and, finally, flux analysis, recently suggested as an '-omics' approach in its own right ('fluxomics' [27]). Table 1 gives a short overview of the typical characteristics of these variants. We are aware that all shades of intermediate analytical set-up might exist and that single analytical technologies, such as GC-MS or LC-MS, might enable all four levels of information to be obtained, depending on the chosen experimental set-up.

Quantification of concentrations predates the '-omic' era and was the first means to characterize the metabolic

Table 1. Overview of the four general variants in the toolbox of metabolome analyses. Properties of fingerprinting, profiling, pool size and flux analysis are described for typical analyses

	Fingerprinting	Profiling	Pool size analysis	Flux analysis
Major field of application	Functional genomics, diagnostics	Functional genomics, molecular physiology	Biochemistry, biotechnology, molecular physiology	Biotechnology, modeling
Major result	Sample classification based on apparent metabolite pattern	Relative quantification of changes in metabolite pool size, identification and discovery of novel metabolites	Absolute quantification of metabolite pools	Quantification of metabolite flux
Sample composition	High complexity (minimal pre-purification)		Low complexity (partial or highly selective purification)	High complexity (minimal prepurification) possible
Sample throughput	High	High-medium	Low (might be extremely high when dedicated to a single metabolite)	Medium–low
Analytical technology	Nonhyphenated technologies possible	Hyphenated technologies required	Combination of hyphenated or nonhyphenated technologies (dependent on the means of prepurification)	
Metabolite coverage	Limited only by choice of metabolite extraction and analytical technology		Preconceived, that is, limited to a predefined set of targeted metabolites	
	Fingerprinting	**Profiling**	**Pool size analysis**	**Flux analysis**
Metabolite identification	Identification of metabolites not required	Identification of as many metabolites as possible	Unambiguous metabolite identification required	Unambiguous metabolite and mass isotopomer identification required
Metabolite concentrations	The concentration of the most abundant metabolite determines the highest possible sample load. The dynamic range of the instrument defines the detection limit of coanalyzed minor metabolites		Prepurification enables concentration of trace metabolites and thus adaptation to the sensitivity range of the analytical instrument. The dynamic range of instrumental analysis is thus nonlimiting.	
Required control experiments	Detector response is corrected for the initial amount of sample and total losses of material during sample preparation and handling	In addition, analysis of recovery, detection limits and linearity of detector response of all known metabolites	In addition, quantitative calibration of the detector response by dilution of a series of pure metabolites	In addition, tracer experiments with radioactive or stable isotope-labeled metabolites
Analytical trade-off	The precision of metabolite identification and quantification is sacrificed for optimised sample through-put.	Absolute quantification is substituted for relative quantification in exchange for full metabolite coverage and medium to high sample throughput	The number of analyzed metabolites is restricted in exchange for precise quantification	The number of analyzed metabolites is restricted in exchange for precise quantification of metabolite mass isotopomers

Box 1. Experimental set-up of mass isotopomer ratio profiling

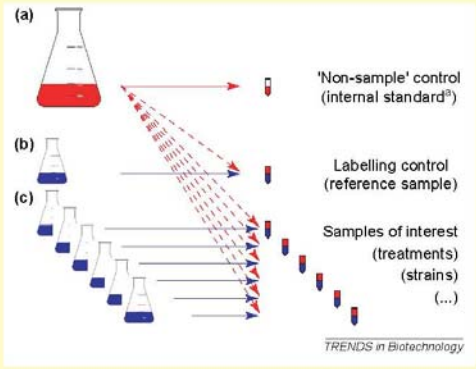


Figure I. (a) A yeast parent strain is grown on pure U-^{13}C-glucose (99 atom %)* in synthetic defined media (red). In **(b)**, an identical culture is prepared with unlabeled glucose (blue). In **(c)**, experiments on different strains or treatments are performed with unlabeled carbon sources (blue). Equal amounts of culture (a) are combined with samples of (b) or (c). Labeled samples serve as analytical internal standards and are typically monitored by 'non-sample' controls. The labeling control (b) checks for inherent changes owing to ^{13}C-labeling (see Box 3b). Relative changes in metabolite pool size are determined by mass isotopomer ratio, as exemplified in Box 3a.*The vitamin and auxotrophic supplements can be non-labelled (Box 2, point 3).

make-up of biological samples via metabolite pool sizes, that is, the metabolic phenotype [2,4]. The availability of pure metabolites is a prerequisite for quantitative analytical methods. Thus, quantification of concentrations might be considered as preconceived: only information on the metabolite under scrutiny is retrieved. The reason for this immanent bias is obvious – quantitative analytics requires calibration of detector signals to metabolite amount. Unequivocal identification and determination of detection limit, linear range, upper loading limit and inherent method variability are only possible when pure reference metabolites are available. In addition, recovery analyses are necessary to check the influence of the sample composition on quantitative detection, the so-called matrix effect [16,28,29].

The discovery of radioactive and stable isotopes and development of specific detectors for radioactive decay and the single unit mass differences of isotopes sparked investigations of metabolite flux [18,19,27,30–35]. Flux analysis is again targeted to the number of preconceived metabolites, and requires application of an isotope tracer that leads to a partial incorporation of isotopes into metabolite pools [30,32–35]. Demands on subsequent analytical resolution are high. Resolution is necessary not only for each metabolite but also for each isotope variant, the mass isotopomers.

By contrast, profiling and fingerprinting technologies are aimed at detection of all metabolites that fall within the range of the chosen technologies [1–16,24,25]. Accordingly, the concept of metabolite purification before analysis is reverted into combinatorial approaches that aim to combine as many metabolites as possible into one analysis. Thus, profiling and fingerprinting could be termed 'non-biased', that is, limited only by the scope of the chosen means of metabolite extraction and analytical technology [4]. The major appeal of these approaches is the potential of discovery. Novel or unexpected metabolites can now be linked to physiological processes or gene function and used as biomarkers [25,26].

The nonbiased approach, however, comes at a cost. Typical fingerprinting analyses use noncalibrated detector readings obtained from complex metabolite mixtures for sample classification and biomarker screening. Analysis is, as a rule, thoroughly checked by nonsample controls and calibrated to reference samples, so that relative changes in signals, as compared with the reference sample, can be calculated. This set-up, although highly efficient in screening and classifying high numbers of samples, could be deemed to be insufficient for four main reasons. Firstly, little or no effort is put into assigning metabolite identity to detector signals. Applications are thus restricted to sample classification, without the potential to unravel the underlying metabolic and physiological cause. Secondly, single metabolites could be represented by multiple detector readings, such as diverse NMR signals or wavelengths. Therefore, it is conceivable that a single or few abundant metabolites might unknowingly dominate sample classification. Thirdly, artifacts caused by laboratory contamination cannot be completely ruled out and will have an impact on fingerprints. Fourthly, and most importantly, metabolite recovery is not controlled in fingerprints. Thus, observed apparent changes do not necessarily reflect direct metabolic changes within the sample but might actually be caused by matrix effects.

For the above reasons, metabolite profiling was suggested; this aims to identify as many metabolites as possible. This concept has become feasible with the advent of hyphenated technologies that enable joined measurement of multiple chemical properties and exploit these properties, for example mass and chromatographic retention, for separation and metabolite identification. This increase in resolution and chemical information represents a substantial improvement because values that were obtained from hyphenated technologies and that describe compound properties that allow metabolite identification can now be exchanged between laboratories. As a testing ground, an open exchange of metabolite identification based on GC-MS mass spectra and chromatographic retention was envisioned [36], has now been initiated (mass spectral libraries at CSB.DB, http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html) and will hopefully be joined by efforts on other technology platforms.

The fundamental advantage of profiling is the opportunity to meet quantitative standards for all identified metabolites within profiles, especially the highly important aspect of metabolite recovery. The technological means for control of matrix effects in metabolite profiles have been suggested previously, namely the use of stable isotope labeled internal standards [3,37]. However, the high costs of chemical synthesis and the apparent lack of availability of standard synthesis for as-yet unidentified metabolites

Box 2. Head to tail comparison of gas chromatography–mass spectrometry (GC-MS) spectra from separate ^{13}C-labeled and ^{12}C-metabolite preparations

In Figure II mass spectra show the number of carbon atoms in all those mass fragments which originate from metabolites.

1: High labeling efficiency is essential because the chances of obtaining a fully labeled mass isotopomer decrease when atom numbers increase (see Glossary). Up to C_{28}, we found unambiguous mass isotopomer distribution in metabolites from yeast grown on pure U-^{13}C-glucose (99 atom %).

2: Incomplete labeling, although insufficient for the determination of high carbon numbers, still enables quantification by mass isotopomer ratios. For high molecular weight metabolites, *in vivo* labelling of less abundant elements, for example N, chemical tagging or analysis of low molecular weight constituents is advisable, such as are employed in proteome analysis [41,42].

3: Addition of unlabeled essential vitamins and auxotrophic supplements to microbial cultures causes respective products to be unlabeled. For example, we found NAD^+ to be fully labeled at the 15 carbon atoms which are ultimately synthesized from glucose. The residual six carbon atoms resulting from the nicotinic acid vitamin supplement were unlabeled.

[a]The GC-MS metabolite profiling requires chemical derivatization by N-methyl-N-(trimethylsilyl)-trifluoroacetamide (MSTFA). This reagent introduces a specific number of trimethylsilyl moieties (TMS) to each metabolite molecule, as is indicated in brackets.

[b]Mass fragments at 73 and 147 mass units are generated exclusively from TMS moieties.

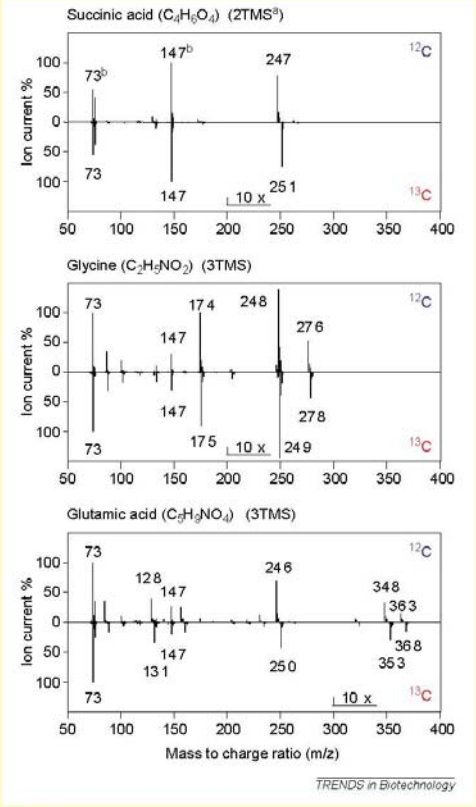


Figure II.

made this suggestion appear unfeasible for general application.

Recent advances in *in vivo* labeling of microorganisms, specifically yeast, that use and modify the experimental concepts of flux analysis, open up novel perspectives for avoiding the analytical pitfalls of metabolite profiling [38].

Quantitative metabolite profiling by mass isotopomer ratios

Efforts towards quantitative profiling technologies were essential for the general acceptance of transcriptomics and proteomics as semiquantitative methods [39–44]. Approaches that introduce a differential label in the course of chemical analysis currently prevail in quantitative transcriptome [39,40] and proteome [41–44] analyses, for example isotope-coded protein-tagging techniques [41,42], protein fluorescence labeling [43,44] and two-color nucleotide labeling by fluorescent probes [39,40]. Although all RNA or protein molecules have common chemical moieties that can be exploited for directed chemical labeling, in metabolome analyses comprehensive chemical tagging technologies are impossible, not least because of the high chemical diversity of metabolites. The most elegant solution to this problem is the introduction of label at atomar level through *in vivo* labeling of the biological reference sample [38]. The two elements found most abundantly in living organisms are carbon and hydrogen ^{13}C carbon can be easily supplied in the form of pure carbon sources to synthetic defined media of microbes or as carbon dioxide to photosynthetic organisms. By contrast, complete hydrogen replacement requires deuterated water and nutrients. For this reason, ^{13}C carbon labeling appears most promising. *In vivo* labeling can be performed in a similar way to a typical flux experiment but must be directed towards complete labeling, for example by feeding pure U-^{13}C-glucose (99 atom %), starting with colony plating (see Figure I in Box 1). Replacement of other elements essential to life

appears feasible but is either experimentally more difficult or restricted to limited parts of the metabolome.

Novel applications of ^{13}C-saturated microbial metabolomes

Several fascinating uses of ^{13}C-saturated microbial metabolomes are envisioned and some have already been pursued. The most essential application is internal standardization of metabolite profiling experiments (see Figure I in Box 1 and Figure III in Box 3) by addition of standardized extracts from ^{13}C-saturated microbial metabolomes. This procedure enables correction for the recovery of each metabolite [38]. Moreover, when the same metabolite is profiled using different MS-based technology platforms, the isotope mass ratio will be identical and independent of suppression effects, as occurs for example in ESI-MS [28] or matrix-assisted laser desorption–time of flight (MALDI-TOF)-MS [29,33] experiments. Thus, isotope mass ratio profiling has the potential to finally unify measurements obtained from the multitude of relevant profiling technologies.

Although this issue alone justifies efforts to establish ^{13}C isotope mass ratio metabolite profiling, having isotope mass ratios at our disposal opens the door for the use of MS-based methods for the purpose of quantification of pool size, which absolutely require standardization by stable isotope techniques, such as MALDI-TOF-MS [29,31,45]. In addition, the enrichment of trace compounds or unstable metabolites that is typically accompanied by high and highly variable metabolite losses is now a feasible procedure in metabolite profiling.

The mass spectra of ^{13}C-labeled molecules provide information concerning the number of carbon atoms in each fragment when mass shifts of highly labeled and unlabeled metabolites are compared (see Figure II in Box 2). This knowledge enhances interpretation of mass spectral fragmentation and elucidation of molecular sum formulas, both essential means to narrow down possible chemical structures of yet unidentified compounds [37].

A variant of typical flux analyses and tracing experiments for pathway identification can be pursued. A stable isotope-labeled metabolome enables analysis of the fate of unlabeled chemicals. Thus, the multitude of cheap and commercially available unlabeled compounds can now be used for tracer and pulse experiments within a ^{13}C-saturated metabolome. This approach appears feasible because we have demonstrated that only minor changes in metabolite levels occur upon ^{13}C labeling (see Figure IIIb in Box 3).

A final, and possibly trivial, but highly effective advantage of *in vivo* stable isotope labeling is the very fact that a metabolite is labeled *in vivo*. This fact is direct proof that the compound is indeed a metabolite and is not one of the possible laboratory contaminants, which have hitherto been tedious to detect and avoid.

Box 3. Quantification by gas chromatography–mass spectrometry mass isotopomer ratio profiling

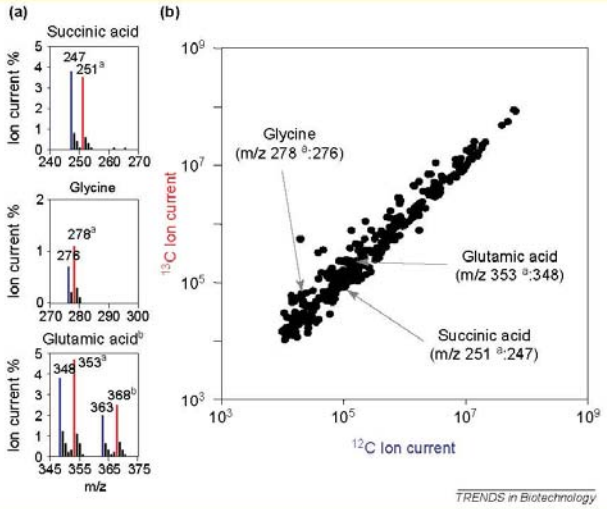


Figure III. **(a)** shows fragment pairs of labeled and unlabeled mass isotopomers representing the same metabolite. Ion currents reflect the relative changes in metabolite abundance. **(b)** Plot of labeled over unlabeled metabolite fragments from a mass isotopomer ratio profile, demonstrating that yeast cultures – in this case overnight batch cultures – exhibit small but perceptible changes in metabolite levels upon *in vivo* ^{13}C labeling (This plot represents the labeling control experiment shown in in Box 1).
[a]Mass fragments which represent the ^{13}C-labeled mass isotopomer, that is, the specific internal standard for this metabolite.
[b]Metabolites can be monitored by one or multiple mass isotopomer pairs for quantification and confirmation.
[c]Labeled mass isotopomers, especially those with fewer than three carbon atoms, are best corrected for natural stable mass isotopes.

In conclusion, we are convinced that mass isotopomer ratio metabolite profiling will not only enhance accurate and quantitative monitoring of the metabolome but also enable comparison of quantitative results from diverse analytical sources and thus take into account the fact that metabolome data need to be generated through a set of diverse analytical platforms.

References

1 Oliver, S.G. *et al.* (1998) Systematic functional analysis of the yeast genome. *Trends Biotechnol.* 16, 373–378
2 Roessner, U. *et al.* (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* 13, 11–29
3 Fiehn, O. *et al.* (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* 18, 1157–1161
4 Fiehn, O. (2002) Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.* 48, 155–171
5 Glassbrook, N. *et al.* (2000) Metabolic profiling on the right path. *Nat. Biotechnol.* 18, 1142–1143
6 Goodacre, R. *et al.* (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.* 22, 245–252
7 Trethewey, R.N. *et al.* (1999) Metabolic profiling: a Rosetta stone for genomics? *Curr. Opin. Plant Biol.* 2, 83–85
8 Nicholson, J.K. *et al.* (1999) 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 29, 1181–1189
9 Mamer, O.A. (1994) Metabolic profiling – a dilemma for mass spectrometry. *Biol. Mass Spectrom.* 23, 535–539
10 Sumner, L.W. *et al.* (2003) Plant metabolomics: large scale phytochemistry in the functional genomics era. *Phytochemistry* 62, 817–836
11 Goodacre, R. *et al.* (2002) Metabolic profiling using direct infusion electrospray ionisation mass spectrometry for the characterisation of olive oils. *Analyst* 127, 1457–1462
12 Castrillo, J.I. *et al.* (2003) An optimized protocol for metabolome analysis in yeast using direct infusion electrospray mass spectrometry. *Phytochemistry* 62, 929–937
13 Huhman, D.V. and Sumner, L.W. (2002) Metabolic profiling of saponins in *Medicago sativa* and *Medicago truncatula* using HPLC coupled to an electrospray ion-trap mass spectrometer. *Phytochemistry* 59, 347–360
14 Soga, T. *et al.* (2002) Simultaneous determination of anionic intermediates for *Bacillus subtilis* metabolic pathways by capillary electrophoresis electrospray ionization mass spectrometry. *Anal. Chem.* 74, 2233–2239
15 Raamsdonk, L.M. *et al.* (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.* 19, 45–50
16 Kopka, J. *et al.* (2004) Metabolite profiling in plant biology: platforms and destinations. *Genome Biol.* 5, 109–117
17 Christensen, B. and Nielsen, J. (1999) Metabolic network analysis: a powerful tool in metabolic engineering. *Adv. Biochem. Eng. Biotechnol.* 66, 209–231
18 Sauer, U. *et al.* (1999) Metabolic flux ratio analysis of genetic and environmental modulations of *Escherichia coli* central carbon metabolism. *J. Bacteriol.* 181, 6679–6688
19 Fischer, E. and Sauer, U. (2003) Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using GC-MS. *Eur. J. Biochem.* 270, 880–891
20 Forster, J. *et al.* (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.* 13, 244–253
21 Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organisation. *Nat. Rev. Genet.* 5, 101–113
22 De Koning, W. and Van Dam, K. (1992) A method for the determination of changes of glycolytic metabolites in yeast on a subsecond time scale using extraction at neutral pH. *Anal. Biochem.* 204, 118–123
23 Buchholz, A. *et al.* (2002) Metabolomics: quantification of intracellular metabolite dynamics. *Biomol. Eng.* 19, 5–15
24 Roessner, U. *et al.* (2000) Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J.* 23, 131–142
25 Harrigan, G.G. and Goodacre, R. eds (2003) *Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis*, Kluwer Academic Publishers
26 Ilyin, S.E. *et al.* (2004) Biomarker discovery and validation: technologies and integrative approaches. *Trends Biotechnol.* 22, 411–416
27 Sauer, U. (2004) High-throughput phenomics: experimental methods for mapping fluxomes. *Curr. Opin. Biotechnol.* 15, 58–63
28 Matuszewski, B.K. *et al.* (2003) Strategies for the assessment of matrix effect in quantitative bioanalytical methods based on HPLC-MS/MS. *Anal. Chem.* 75, 3019–3030
29 Guo, Z. *et al.* (2002) A method for the analysis of low-mass molecules by MALDI-TOF mass spectrometry. *Anal. Chem.* 74, 1637–1641
30 Szyperski, T. (1998) ^{13}C-NMR, MS and metabolic flux balancing in biotechnology research. *Q. Rev. Biophys.* 31, 41–106
31 Wittmann, C. and Heinzle, E. (2001) Application of MALDI-TOF MS to lysine-producing *Corynebacterium glutamicum*. A novel approach for metabolic flux analysis. *Eur. J. Biochem.* 268, 2441–2455
32 Wiechert, W. (2001) ^{13}C metabolic flux analysis. *Metab. Eng.* 3, 195–206
33 Wittmann, C. (2002) Metabolic flux analysis using mass spectrometry. *Adv. Biochem. Eng. Biotechnol.* 74, 39–64
34 Hellerstein, M.K. (2003) *In vivo* measurement of fluxes through metabolic pathways: the missing link in functional genomics and pharmaceutical research. *Annu. Rev. Nutr.* 23, 379–402
35 Nielsen, J. (2003) It is all about metabolic fluxes. *J. Bacteriol.* 185, 7031–7035
36 Wagner, C. *et al.* (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry* 62, 887–900
37 Fiehn, O. *et al.* (2000) Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal. Chem.* 72, 3573–3580
38 Mashego, M.R. *et al.* (2004) MIRACLE: mass isotopomer ratio analysis of U-^{13}C-labeled extracts. A new method for accurate quantification of changes in concentrations of intracellular metabolites. *Biotechnol. Bioeng.* 85, 620–628
39 Schena, M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470
40 Lockhart, D.J. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14, 1675–1680
41 Gygi, S.P. *et al.* (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 17, 994–999
42 Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* 422, 198–207
43 Unlu, M. *et al.* (1997) Difference gel electrophoresis: a single method for detecting changes in protein extracts. *Electrophoresis* 18, 2071–2077
44 Van den Bergh, G. *et al.* (2004) Fluorescent two-dimensional difference gel electrophoresis unveils the potential of gel based proteomics. *Curr. Opin. Biotechnol.* 15, 38–43
45 Kang, M-J. *et al.* (2001) Application of automated matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry for the measurement of enzyme activities. *Rapid Commun. Mass Spectrom.* 15, 1327–1333
46 Platzner, I.T. (1997) *Modern Isotope Ratio Mass Spectrometry*, John Wiley & Sons Inc

www.sciencedirect.com

Available online at www.sciencedirect.com
ScienceDirect
Phytochemistry 68 (2007) 2258–2272
PHYTOCHEMISTRY
www.elsevier.com/locate/phytochem

GC-EI-TOF-MS analysis of *in vivo* carbon-partitioning into soluble metabolite pools of higher plants by monitoring isotope dilution after $^{13}CO_2$ labelling

Jan Huege, Ronan Sulpice, Yves Gibon, Jan Lisec, Karin Koehl, Joachim Kopka *

Max Planck Institute of Molecular Plant Physiology, Am Muehlenberg 1, 14476 Potsdam-Golm, Germany

Received 12 January 2007; received in revised form 12 March 2007
Available online 1 May 2007

Abstract

The established GC-EI-TOF-MS method for the profiling of soluble polar metabolites from plant tissue was employed for the kinetic metabolic phenotyping of higher plants. Approximately 100 typical GC-EI-MS mass fragments of trimethylsilylated and methoxyaminated metabolite derivatives were structurally interpreted for mass isotopomer analysis, thus enabling the kinetic study of identified metabolites as well as the so-called functional group monitoring of yet non-identified metabolites. The monitoring of isotope dilution after $^{13}CO_2$ labelling was optimized using *Arabidopsis thaliana* Col-0 or *Oryza sativa* IR57111 plants, which were maximally labelled with ^{13}C. Carbon isotope dilution was evaluated for short (2 h) and long-term (3 days) kinetic measurements of metabolite pools in root and shoots. Both approaches were shown to enable the characterization of metabolite specific partitioning processes and kinetics. Simplifying data reduction schemes comprising calculation of ^{13}C-enrichment from mass isotopomer distributions and of initial ^{13}C-dilution rates were employed. Metabolites exhibited a highly diverse range of metabolite and organ specific half-life of ^{13}C-label in their respective pools (^{13}C-half-life). This observation implied the setting of metabolite specific periods for optimal kinetic monitoring. A current experimental design for the kinetic metabolic phenotyping of higher plants is proposed.

Keywords: *Arabidopsis thaliana* Col-0; *Oryza sativa* IR57111; ^{13}C-carbon; $^{13}CO_2$-carbondioxide; Dynamic flux analysis; Electron impact ionization (EI); Gas chromatography (GC); Metabolite profiling; Stable isotope dilution; Time-of-flight mass spectrometry (TOF-MS)

1. Introduction

The labelling of plants with CO_2 using both radioactive (e.g. Calvin, 1956, 1964) or stable isotope tracing (e.g. Schaefer et al., 1975, 1980; MacLeod et al., 2001; Schwender et al., 2004) has been used since decades utilizing the main entry points of CO_2 into plant metabolism, namely ribulose-1,5-diphosphate carboxylase (EC 4.1.1.39) and phosphoenolpyruvate carboxylase (EC 4.1.1.31). Over time, CO_2 tracing yielded ground-breaking biological insights into photosynthetic carbon assimilation, photorespiration and metabolism and, thus, into essential life-sustaining physiological mechanisms on earth.

With the increasing availability of stable isotopes, methods for the *in vivo* labelling of plants are conceivable. Successful methods for the complete and saturating ^{13}C-labelling were reported previously using a microbial model, such as *Saccharomyces cerevisiae* (Birkemeyer et al., 2005). In this work we present and characterize a method for the full *in vivo* ^{13}C-labelling of higher plants.

Abbreviations: EI, electron impact ionization; GC, gas chromatography; MS, mass spectrometry; MST, mass spectral tag; RI, retention time index; TOF, time-of-flight.

* Corresponding author. Tel.: +49 331 567 8262; fax: +49 331 567 898262.
E-mail address: Kopka@mpimp-golm.mpg.de (J. Kopka).

0031-9422/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.phytochem.2007.03.026

With this prerequisite in place isotope dilution after $^{13}CO_2$ labelling was monitored. The kinetic measurement of isotope decay from ^{13}C to ^{12}C can be performed under ambient atmospheric conditions, thus allowing sampling with minimal experimental disturbance. As a consequence experiments with extended kinetic monitoring of CO_2-dilution under diverse regimes of environmental conditions are now conceivable. Thus we aim to contribute to the ongoing discussion and development of empirical flux estimations in the field of plant physiology (e.g. Fernie et al., 2005; Baxter et al., 2007). We specifically intend to work towards dynamic flux estimations (Ratcliffe and Shachar-Hill, 2006) as a tool for the phenotypic analysis of gene function in higher plants.

Besides nuclear magnetic resonance (NMR), mass spectrometric (MS) analysis has traditionally been used for flux analyses, for example MALDI-TOF (e.g. Wittmann and Heinzle, 2001), quadrupole based GC–MS (e.g. Dauner and Sauer, 2000), ion trap mass spectrometry (e.g. Klapa et al., 2003) or LC triple quadrupole mass spectrometry (e.g. van Winden et al., 2005). We employed the widely applied gas chromatography electron impact ionization time-of-flight mass spectrometry, in short GC-EI-TOF-MS technology, for metabolite profiling (e.g. Wagner et al., 2003; Lisec et al., 2006; Erban et al., 2006). GC-EI-TOF-MS is a non-scanning mass spectrometric technology, which allows simultaneous monitoring of a mass range and high acquisition rates of 10–500 mass spectra s^{-1}. Thus apparent fragment ratios are not subject to artefacts caused by the temporal offset of the sequential mass recording, which is inherent to mass scanning technologies, such as the quadrupole or ion trap GC–MS. The feasibility of mass isotopomer monitoring of methoxyaminated and silylated derivatives by GC–MS metabolite profiling has been demonstrated earlier (Roessner-Tunali et al., 2004; Baxter et al., 2007). The decision to use this specific mode of derivatization and analytical monitoring was made in view of the high synergy to be expected of method development and metabolite identification efforts in the metabolite profiling field (e.g. Schauer et al., 2005a). Substantial instrumental progress has also been made, for example by GC×GC-TOF-MS (e.g. Sinha et al., 2004a,b; Kell et al., 2005) implementation. Furthermore, a highly versatile tool box of metabolite fractionation and chemical derivatization schemes awaits exploration (Kopka, 2006a).

In the following study we perform a technological assessment of the combination of mass isotopomer analysis using the GC-EI-TOF-MS profiling method and monitoring of isotope dilution after $^{13}CO_2$ labelling. We specifically address the use of populations of replicate, genetically identical plants for flux studies and describe fundamental technological requirements. First results are presented, which demonstrate the feasibility and potential but also the current limitations of this novel combination of techniques.

2. Results and discussion

2.1. GC-EI-MS fragmentation analysis

The electron impact (EI) fragmentation pattern of trimethylsilylated and methoxiaminated metabolite derivatives (analytes), which are observed in routine metabolite profiles (e.g. Erban et al., 2006; Fiehn et al., 2000a; Lisec et al., 2006; Roessner et al., 2000) delimit the potential of this profiling method for the multi-parallel analysis of metabolic fluxes using ^{13}C-stable isotope dilution. Specifically the targeted retrieval of quantitative mass isotopomer information and the calculation of isotope enrichment from mass isotopomer distributions require the thorough interpretation of GC-EI fragmentation patterns and the knowledge of the sum formula of each analyzed mass fragment. As suggested previously (Birkemeyer et al., 2005), we used mass spectral tags (MSTs; cf. the definition made by Desbrosses et al., 2005 refined by Kopka, 2006b) with ambient isotopic composition and MSTs from *in vivo* ^{13}C-labelled material for the interpretation of mass spectral fragmentation patterns (Fig. 1). This interpretation effort was based on the detailed EI fragmentation patterns of trimethylsilylated and methoxiaminated carbohydrates (DeJongh et al., 1969; Laine and Sweeley, 1973; MacLeod et al., 2001; Sanz et al., 2002) or amino acids (Abramson et al., 1974; Bergström et al., 1970; Leimer et al., 1977), which have been previously published. Furthermore, interpretation was supported by the mass shifts observed upon *in vivo* ^{13}C-labelling (Fig. 1).

The fragmentation patterns of 58 analytes, comprising approximately 100 electron impact fragments, were analyzed. Representative sugars, organic acids, amino acids, amines and polyols were chosen (cf. Supplementary file 1). This selection represents approximately 7.6% of the current non-redundant Golm Metabolome Database compendium (GMD; Kopka et al., 2005; Schauer et al., 2005a; http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html). The sum formula of each mass fragment was deduced from known general fragmentation reactions and available references (see above). For further referencing all obtained fragment information was linked to the mass spectrum identifier system used by GMD (MPIMP-ID). The ^{12}C- and ^{13}C-monoisotopic masses and the respective number of carbon atoms, which originate from the carbon-backbone of each metabolite and are not introduced by chemical derivatization reagents, were empirically determined from ambient and fully *in vivo* labelled MSTs (Fig. 1). In agreement with the GC-EI-TOF-MS instrument specifications, monoisotopic masses are given at full mass unit precision.

Mass fragments comprising the full metabolite carbon-backbone, such as M^+ (molecular ion) or $M-15^+$ (i.e. a mass fragment generated from M through a loss of 15 a.m.u.), were mostly present at low intensities and in some cases below detection limit. Through our effort we now offer alternative fragment ions for mass isotopomer

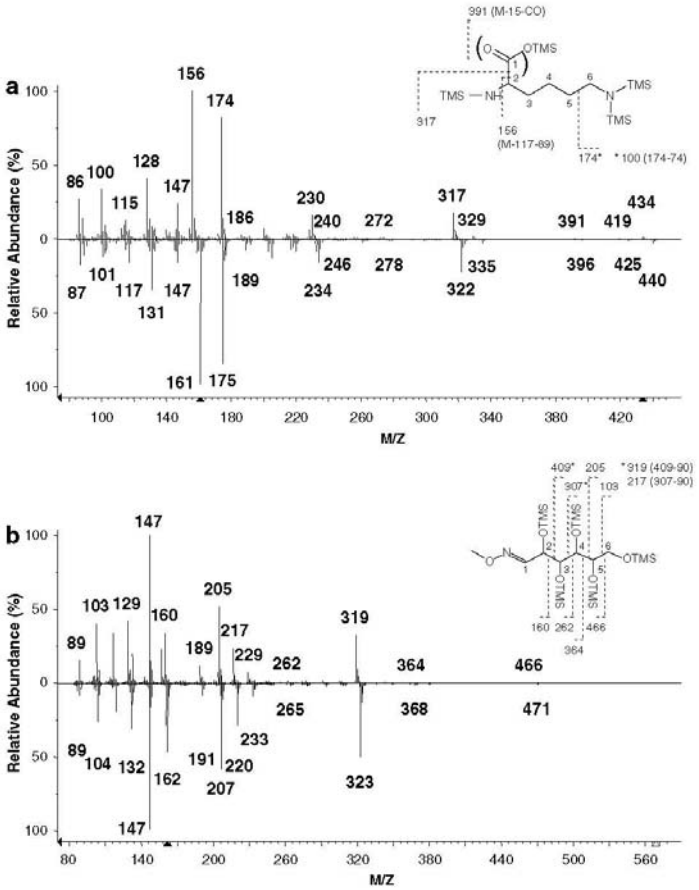


Fig. 1. Representative interpretation of GC-EI-TOF-MS mass spectra of (a) a lysine (4TMS), A192003*, and (b) a hexoaldose analyte, for example glucose (1MEOX) (5TMS), A191001*. Availability of ambient and fully labelled ^{13}C-mass isotopomer specta supported interpretation of known general fragmentation reactions from previous reports. Note (1) the benefit of the methoxyamine chemical tag, which allowed distinction of a $C_1 \rightarrow C_6$ and a $C_6 \rightarrow C_1$ fragment series of reducing sugars. (2) Some fragmentations may be indicative of specific functional moieties and thus useful for functional group mass spectrometry, such as the occurrence of m/z 174 → 175, which represents $C_7H_{20}NSi_2^+$, typical of primary amines, or of m/z 218 → 220 representing $C_8H_{20}NO_2Si_2^+$ with $C_{1,2}$ of amino acids. * Mass spectrum identifier, MPIMP-ID, as used by the Golm Metabolome Database (GMD; http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html, cf. fragmentation details reported within Supplementary file 1). Mass spectra were scaled to the fragment ions, (a) m/z 156 → 161 or (b) m/z 147, and are displayed head to tail.

analysis. As described previously *tert*-butyldimethylsilylation (e.g. Dauner and Sauer, 2000; Fiehn et al., 2000b; Birkemeyer et al., 2003; Klapa et al., 2003) was successfully assessed as an alternative chemical derivatization strategy to obtain full carbon-backbone information through typically abundant M^+, $M-15^+$, as well as $M-57^+$ fragments. This derivatization method, while perfectly suited for complementary analyses of amines, organic acids and amino acids, was, however, not advisable for the analysis of compounds with multiple vicinal diols, for example saccharides. This derivatization scheme was not further used in this study, because we primarily aimed at highest possible comprehensiveness.

Ample information on mass fragments representing different metabolite substructures was available (Fig. 1). Analytes typically exhibited more than one mass fragment, which was amenable to ^{13}C-isotope dilution analysis. However, in agreement with most MS-based methods of flux estimation, the available substructure information rarely provided sufficient information to extract exact labelling information of each single carbon atom. In contrast multiple mass fragments were typically available, which

characterized complementary scission products or partially overlapping metabolite substructures. In some cases fragmentation reactions allowed assessment of the same metabolite substructure using either different mass fragments of the same analyte, for example *m/z* 307 and the daughter product *m/z* 217, which is generated by neutral loss of C_3H_9SiOH (e.g. Fig. 1b). In other cases mass fragments of alternative derivatization products can be used, for example the two *E/Z*-methoxyamination products of reducing sugars or different silylation products, such as the glycine (2TMS) and glycine (3TMS) products (cf. the information on fragmentation sequence and alternative derivatization products within Supplementary file 1).

Cases of alternative sum formula interpretations could be solved empirically by the ^{13}C-induced mass shift. When possible, we annotated the carbon atoms of the metabolite, which were represented by the observed mass fragments (Supplementary file 1). Furthermore, we indicated those mass fragments, which may originate from different substructures of the same compound. Symmetrical compounds were typical for generating ambiguous substructure information, as were oligomers, such as di- or trisaccharides. Other cases of potentially "non-unique" origin, also indicated in Supplementary file 1, can only be recognized and verified after detailed analysis of positional labelled metabolites (e.g. MacLeod et al., 2001).

Methoxyamination of carbonyl-moieties proved to be highly efficient in simplifying the fragmentation pattern of sugar derivatives and allowed discrimination of two fragmentation series, containing the reduced or non-reduced part of the molecules, respectively (Fig. 1b). Many mass fragments were typical of compound classes. For example *m/z* 160 and 262 are indicative of aldoses. These mass fragments represent $C_{1,2}$ and $C_{1,2,3}$ of the sugar, respectively (Fig. 1a). Amino acids and amines also yielded examples of fragments, which represent specific functional groups. For example, the mass fragment *m/z* 174 shifted to 175 when ^{13}C-labelled. This fragment represents the typical ion $C_7H_{20}NSi_2^+$, which contains C_1 of primary amines. Also the mass fragment *m/z* 218 shifted to 220 when ^{13}C-

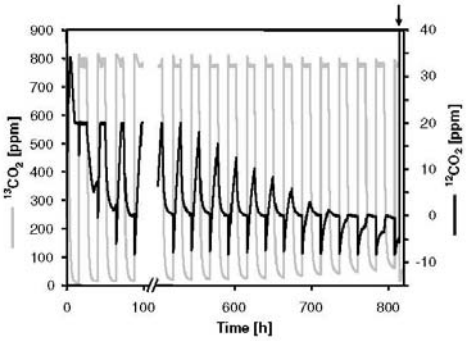


Fig. 2. Saturating $^{13}CO_2$ labelling regime of a population of 60 *Arabidopsis thaliana* Col-0 plants grown in a 6 L hydroponic culture. Plantlets were pre-grown in ambient conditions. At the four-leaf stage a developmentally homogenous population was transferred to $^{13}CO_2$ labelling. The upper $^{12}CO_2$ threshold was set to ~20 ppm during the day; above threshold emitted $^{12}CO_2$ was replaced by $^{13}CO_2$. At night the atmosphere was kept CO_2-free to save $^{13}CO_2$. Growth conditions were a 10 h day at 113.5 (SD = 6.0) μmol m^{-2} s^{-1}, 23.7 (SD = 0.5) °C, 71.8 (SD = 2.4)% relative humidity, 22.3 (SD = 0.3)% O_2, 776.6 (SD = 78.2) ppm total $^{13}CO_2$ and 14 h night at 6.8 (SD = 10.4) μmol m^{-2} s^{-1}, 19.7 (SD = 1.1) °C, 68.7 (SD = 2.0)% relative humidity, 22.3 (SD = 0.3)% O_2, 101.9 (SD = 154.1) ppm total $^{13}CO_2$. In this experiment 1.2 L $^{13}CO_2$ was spent to generate 1.528 g (FW) shoot and 1.796 g (FW) root material. The arrow indicates the start of the isotope dilution experiment. SD indicates standard deviation.

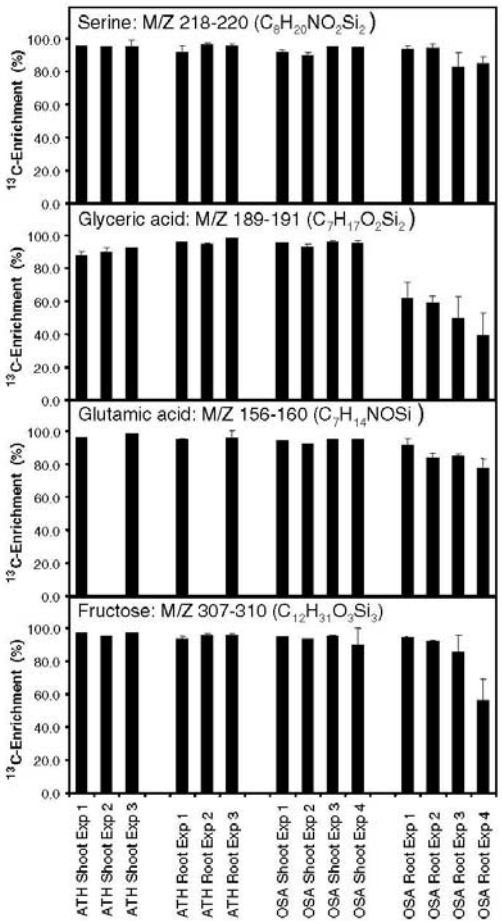


Fig. 3. *In vivo* saturated ^{13}C-isotope enrichment of root and shoot from *Arabidopsis thaliana* Col-0 or *Oryza sativa* plants. Repeatability of serine (A138001), glyceric acid (A135003), glutamic acid (A163001) and fructose (A187002) labelling is demonstrated at the endpoint of ^{13}C-labelling as indicated in Fig. 2. Error bars indicate standard deviation (SD).

labelled. This fragment comprises $C_{1,2}$ of amino acids, namely $C_8H_{20}NO_2Si_2^+$.

The strong EI-induced mass spectral fragmentation of GC-EI analysis may be seen as a drawback for MS-based flux analyses. However, we demonstrated the potential of a detailed analysis of substructures of representative identified metabolites within routine GC-EI-MS profiles. The availability of functional group mass spectrometry (Abramson et al., 1974) ultimately facilitates the first access to the flux analysis of still unknown constituents of metabolite profiles.

The complexity of typical metabolite profiles and the requirement for analysis of multiple mass isotopomers and fragmentation patterns currently limit flux analysis to a subset of the total information, which is routinely available from GC–MS based metabolite profiles. However, the recent introduction of two-dimensional GCxGC-EI-TOF-MS instruments (e.g. Sinha et al., 2004; Kell et al., 2005) will increase the purity of fragment information and thus will extend the application of the method, which is currently implemented and developed in our laboratory using conventional GC-EI-TOF-MS.

2.2. $^{13}CO_2$-labelling of populations of replicate plants

In contrast to microbial flux analyses, which are typically performed in continuous or batch-grown cell cultures, flux studies of higher plants necessitate the use of populations of genetically identical replicated plants. The alternative of subsequent sampling from the same plant bears the risk of concurrent non-controlled wounding reactions and developmental or positional effects. However, monitoring isotope dilution after full $^{13}CO_2$ labelling of the metabolome, as proposed (Birkemeyer et al., 2005), may introduce an additional risk of analytical variability, when compared to conventional flux studies. Therefore, we studied and optimized the ^{13}C-saturated labelling efficiency of metabolite pools. We determined the technical and, even more importantly, the biological plant to plant reproducibility of $^{13}CO_2$-labelling within replicate populations of a hydroponically grown monocotyledon, *Oryza sativa*, and a dicotyledon, *Arabidopsis thaliana*, model species.

2.2.1. Efficiency of $^{13}CO_2$-labelling

The final labelling efficiency per se may not directly influence the accuracy of flux determinations. However, highly repeatable, homogenous labelling of all metabolite pools, ideally in both root and shoot tissue may be deemed prerequisite for optimal and routine experimental reproducibility of studies which exploit the monitoring of isotope dilution after $^{13}CO_2$ labelling. We used the emission of $^{12}CO_2$ during the photoperiod as an indicator for the metabolic ^{13}C-saturation of the plant populations under investigation (Fig. 2). This representative experiment comprised 60 *A. thaliana* Col-0 plants grown in a 6 L hydroponic culture using an airtight and atmospherically controlled growth chamber. Plantlets were at the four-leaf stage when transferred to $^{13}CO_2$-labelling in this enclosed environment. During the initial days plantlets as well as liquid medium contributed to $^{12}CO_2$ generation. To gradually exclude ^{12}C from the system, CO_2 was removed from the enclosed atmosphere using a CO_2 absorber. During the light phase CO_2 was removed and substituted by $^{13}CO_2$-gas, when $^{12}CO_2$ partial pressure increased above 20 ppm. During night CO_2 was completely removed from the atmosphere. The time required for the $^{12}CO_2$ partial pressure to permanently drop below the 20 ppm threshold and finally below detection limit was dependent on the number of plants, developmental stage and species. In the shown *A. thaliana* experiment (Fig. 2) 1.2 L $^{13}CO_2$ was spent to obtain a total of 1.528 g (FW) shoot and 1.796 g (FW) root material. The average ^{13}C-enrichment ± standard deviation (SD), as reproduced in 3 (*A. thaliana*) and 4 (*O. sativa*) independent labelling experiments was 90.2 (±7.3)% and 78.6 (±15.6)% for shoots and roots of *O. sativa* and 91.5 (±10.5)% and 90.2 (±9.7)% for shoots and roots of *A. thaliana*, respectively.

Details of the behaviour of specific mass fragments and best labelling results, approximately 98% (*A. thaliana*) and

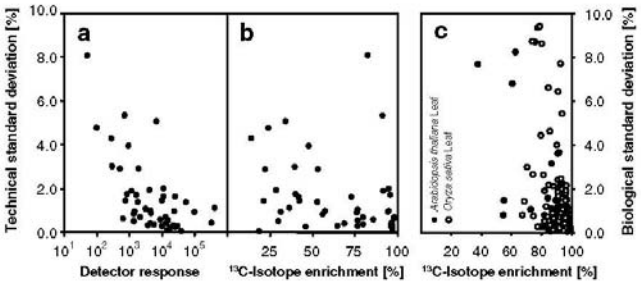


Fig. 4. Influence of (a) detector response, represented as arbitrary ion current, and of (b) ^{13}C-isotope enrichment on the technical standard deviation of ^{13}C-isotope enrichment from complex GC-EI-TOF-MS metabolite profiles of a pool of partially labelled *Arabidopsis thaliana* Col-0 plants. The average standard deviation of 47 representative mass fragments was 1.5% and may increase slightly, when approaching GC-EI-TOF-MS detection limit. (c) The plant to plant standard deviation within a population of simultaneously labelled *Arabidopsis thaliana* Col-0 and *Oryza sativa* plants. Average standard deviation was 1.0% and 1.8%, respectively ($n = 4$ 8).

97% (*O. sativa*), may be found in Supplementary file 1. In general the final labelling efficiency was organ and metabolite specific (Fig. 3). The higher efficiency and reproducibility of labelling obtained with *A. thaliana*, especially in roots, is likely caused by the smaller carbon resources of its seed compared to the large carbon resources of the caryopsis of *O. sativa*. Thus, in order to optimize results with *O. sativa*, the residual endosperm was removed from young seedlings. This finding indicates that plant species with large carbon storage capacities within the seed will be less amenable to analysis of isotope dilution after $^{13}CO_2$ labelling, unless the storage organ can be severed early in seedling development or labelled seeds are used. Besides glyceric acid, and fructose (Fig. 3), glycine and trehalose were the most sensitive indicators for incomplete labelling of the root metabolome, in both species.

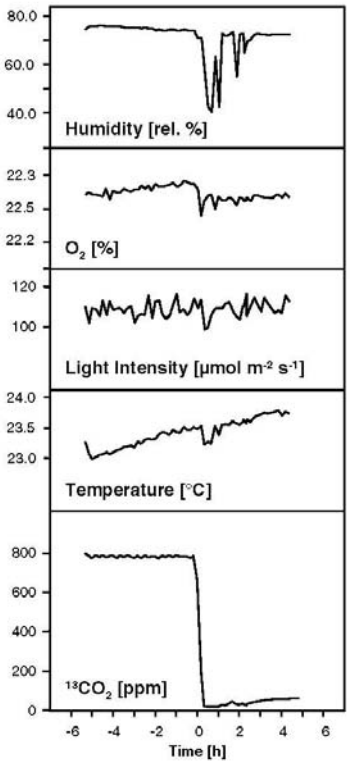


Fig. 5. Homeostasis of environmental conditions during the photoperiod following $^{13}CO_2$ to ambient $^{12}CO_2$ replacement. The parameters were recorded under the conditions of rapid plant sampling, which are required for kinetic studies. Light intensity, temperature and O_2 concentration showed negligible fluctuations. The relative humidity responded to each sampling and the mandatory initial $^{13}CO_2$ flush-out of the growth chamber.

2.2.2. Homogeneity of $^{13}CO_2$-labelling

The technical error of the determination of ^{13}C-labelling efficiency of metabolite pools, which is inherent to GC-EI-TOF-MS metabolite profiling, was estimated using repeated analyses of the same sample. In order to monitor the broad range of possible labelling efficiencies, the sample was prepared from a pool of partially labelled *A. thaliana* Col-0 plants. Average ^{13}C-enrichment (%) and standard deviation was calculated using approximately 50 mass fragments (Fig. 4a and b). The typical technical standard deviation was 1.8% ^{13}C-enrichment. This technical standard deviation was independent of the degree of ^{13}C-enrichment (Fig. 4b), but clearly dependent on the magnitude of the

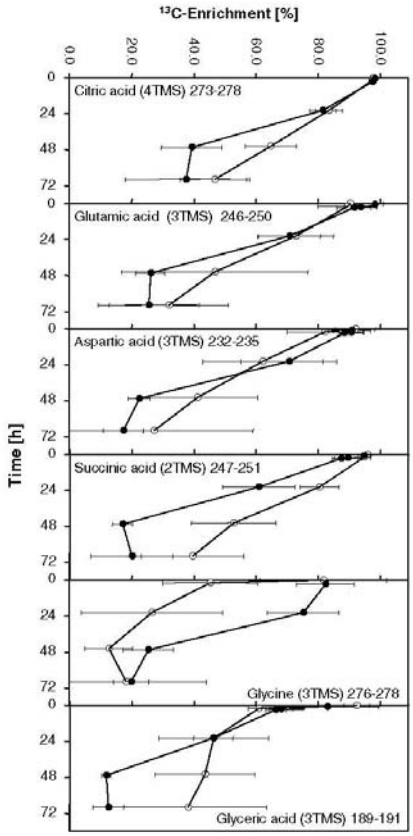


Fig. 6. Long-term kinetics of isotope dilution after $^{13}CO_2$ labelling. Exemplary metabolite pools from *Arabidopsis thaliana* Col-0 shoot (open circles) and root tissue (closed circles) were monitored during a 3 day period following exposure to ambient CO_2. Single plants were analyzed. Plant to plant variation increases dramatically after the first night. Metabolite specific differential behaviour reflecting the shoot to root differences of carbon-partitioning was demonstrated (cf. Table 2, Exp. 1). Error bars indicate standard deviation (SD).

monitored detector response (Fig. 4a). The detection limit of the GC-EI-TOF-MS system used in this study is typically reached at 250–500 arbitrary ion current units with the signal to noise cut-off set to approximately 10.

The "biological" standard deviation calculated from single plant measurements of a simultaneously labelled, genetically homogenous plant population was clearly within the limits of the technical error (Fig. 4c). This result was obtained using shoot material of optimally labelled plant populations. The average standard deviation of all analyzed, maximally labelled mass fragments was 1.0% in shoots and, 2.7% in roots of *A. thaliana*, 1.8% in shoots and 9.0% in roots of *O. sativa*, respectively (Fig. 3 and Supplementary file 1). In conclusion, the full labelling of plants is a prerequisite for optimum experimental homogeneity. Labelling of the root system is subject to higher variability, especially in *O. sativa*. Because *A. thaliana* populations were more amenable to repeatable and homogenous labelling in both organs, we focused on this model in subsequent experiments.

2.3. Assessment of experimental homeostasis during isotope dilution after $^{13}CO_2$ labelling

In plants, the assumption of steady state will be restricted to rare cases. Indeed non-steady state conditions were clearly demonstrated, for example by phenotyping transcriptional, enzymatic, and metabolic changes throughout the diurnal cycle of *A. thaliana* (Blaesing et al., 2005; Gibon et al., 2006). In particular, the first hours after dawn of each photoperiod were shown to yield major changes at all system levels. For this reason we chose to start flux studies in the middle of the photoperiod (Fig. 5). Among a set of selected physico-chemical parameters only the relative humidity could not be kept constant during the initial period of $^{13}CO_2$–$^{12}CO_2$ exchange. All parameters were recorded under conditions of concurrent rapid sampling (Fig. 5). This finding held also true during three subsequent days of extended sampling (data not shown).

The maximum monitoring period, 3 days, was chosen to cover the time required for all metabolite pools to drop below approximately 50% ^{13}C-labelling (Fig. 6). Selected enzyme activities were employed to test for an effect of ^{13}C–^{12}C-substitution during that period. We did not detect a significant impact of ^{12}C-enrichment on enzyme activities during short and extended monitoring periods (Table 1). This might be due to the low protein turnover as compared to metabolic turnover and/or by the fact that ^{13}C-labelled enzymes might retain the same enzymatic properties as ^{12}C-labelled enzymes. We interpret our finding as a first indication that the monitoring of isotope dilution after $^{13}CO_2$ labelling is not impaired by major changes in enzyme kinetic properties.

2.4. Long-term isotope dilution experiments

Metabolite pools have highly diverse ^{13}C-half-life and kinetic behaviour (Fig. 6). The characteristics of isotope dilution kinetics during an extended monitoring period demonstrated clear differences between the metabolite pools of root compared to shoot (Fig. 6). As expected, metabolites linked to respiratory processes, such as citric, malic (data not shown), and succinic acid exhibited a stronger long-term carbon-partitioning into roots. Surprisingly, glyceric acid exhibited a similar pattern, suggesting that this metabolite was also actively metabolized in roots. In contrast, shoots exhibited a preferential carbon-partitioning into glycine and serine (data no shown), most likely due to photorespiration. The long-term kinetic behaviour of isotope dilution in most root pools was equal or even faster compared to shoot pools. Aspartic acid and glycine pools, exhibited a perceptible delay compared to the kinetic behaviour observed in leaves. In some cases, the time course of isotope dilution was found to be biphasic. For example, during the first photoperiod, the incorporation

Table 1
Effect of ambient CO_2 incorporation into ^{13}C-saturated *Arabidopsis thaliana* Col-0 plants on selected enzyme activities

Enzyme		Experiment 1			Experiment 2					
		Day 0			Day 0 [+1.5 h]		Day 2 [±3.5 h]		Day 3 [±3.5 h]	
		Activity [nmol min^{-1} g^{-1} (FW)]	SD	*n*	Fold change	*p*	Fold change	*p*	Fold change	*p*
Isocitrate dehydrogenase (NAD+)	EC 1.1.1.41	88.9	18.5	9	1.00	0.981	0.88	0.441	0.71	0.058
Glycerate kinase	EC 2.7.1.31	2399.5	1576.6	10	1.05	0.898	1.07	0.831	1.27	0.493
Phosphoglycerate kinase	EC 2.7.2.3	37824.9	19332.0	10	1.18	0.520	1.14	0.581	1.22	0.420
Transketolase	EC 2.2.1.1	2186.7	839.9	10	1.25	0.264	1.18	0.408	1.18	0.452

Pools of rosette leaves (shoot) were exposed to ambient CO_2 at 4 h after dawn and sampled between 0 and 1.5 h on day 0 or ±3.5 h on subsequent days. Exemplary enzyme activities were analyzed according to Gibon et al. (2004). Activities of isocitrate dehydrogenase (NAD$^+$) EC 1.1.1.41, glycerate kinase EC 2.7.1.31, phosphoglycerate kinase EC 2.7.2.3, and transketolase EC 2.2.1.1 did not exhibit significant changes upon *in vivo* ^{12}C ^{13}C-isotope exchange neither between independent experiments nor up to 3 days after exposure to ambient CO_2.

into both aspartic and glyceric acid was initially fast, but then slowed down dramatically (Fig. 6).

Not unexpectedly we observed an increasing experimental variability within populations of replicate plants from day to day (Fig. 6). We subsequently focused on the first two photoperiods for a better resolution of the initial dilution kinetics and an improved characterization of the increasing plant to plant variability.

2.5. Short-term isotope dilution experiments

Short-term isotope dilution confirmed the high diversity of ^{13}C-half-life observed in different metabolite pools. In addition, the initial dilution kinetic was apparently linear and shifted subsequently to an asymptotic behaviour (Fig. 7). Unexpectedly, the asymptotic ^{13}C-decay was metabolite dependent and none of the metabolite pools approximated ambient ^{13}C-enrichment. Two parallel processes may apply. (1) Newly assimilated carbon competes with carbon derived from internal stores and an equilibrium phase is reached after the initial linear dilution phase. In other words metabolite pools may reach a quasi steady state after different times. (2) Additionally, in plant cells most metabolites are present in two or more pools, which exhibit slow exchange rates. Furthermore one of these pools would be required to behave metabolically inert, for example the vacuole or possibly the apoplast, whereas the others may represent metabolically highly active pools, such as plastid, mitochondria or cytosol.

Despite the high variability observed during extended monitoring, we were able to accurately monitor isotope dilution by sampling single plants in 1–5 min intervals during the first photoperiod (Fig. 7). Thus initial dilution kinetics can be obtained for metabolites with a fast ^{13}C-half-life, like glyceric acid, sucrose, and fructose. However, sampling needs to remain equally spaced throughout the whole monitoring period because of the high diversity of metabolite ^{13}C-half-life (cf. sucrose, succinic or fumaric acid in Fig. 7). For improved monitoring of metabolites with low ^{13}C-half-life, extension of measurements towards the end of the first photoperiod is advised. Extended monitoring into the second photoperiod showed drastic increase of plant to plant variability. This effect subsequently superseded and thus obscured the kinetic behaviour (insert of Fig. 7). Therefore extension towards 1 day monitoring may only be done for metabolite pools with extremely high ^{13}C-half-life.

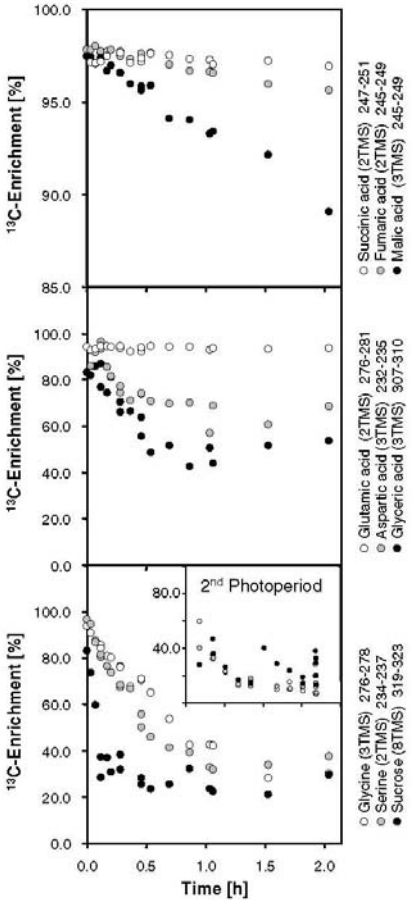


Fig. 7. The short-term kinetic behaviour of ^{13}C-isotope dilution into exemplary metabolite pools of *Arabidopsis thaliana* Col-0 shoot upon exposure to ambient CO_2 (cf. Table 2, Exp. 4). ^{13}C-decay was traced within the initial 2 h of the first light period and throughout the second photoperiod (insert).

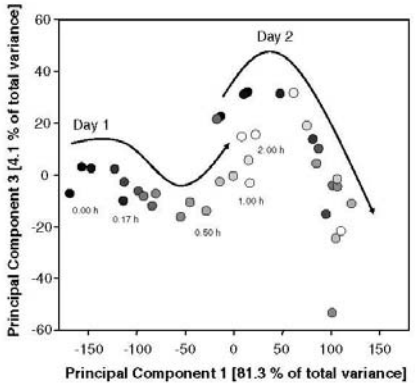


Fig. 8. Principal component analysis of ^{13}C-isotope enrichment kinetics of metabolite pools from *Arabidopsis thaliana* Col-0 shoot upon exposure to ambient CO_2. The temporal sequence of sampling is indicated per day by grey scale. A subset of the metabolites, which show high contribution to the first component of this analysis, is shown in Fig. 7.

2.6. Analysis of the ^{13}C-isotope decay phenotype

We applied principal component analysis (PCA) to analyse the increase in plant to plant variability between the first and second photoperiod (Fig. 8). We focussed on shoot samples of *A. thaliana* and thus avoided obvious variances caused by species and organ differences. Accessing the major variance from a 28 h time course of experiment 4 (Fig. 7 and Table 2), we tested, whether plants of the second photoperiod might follow a common kinetic behaviour despite high apparent variability. All samples appeared to follow the same isotope dilution kinetic during the first and the second day, as was indicated by the first principal component, which comprised 81.3% of the total variance. Interestingly, samples received a slight off-set during the intervening night period. We reason that this observation can be explained by utilization of stored ^{13}C during the night. A non-linear behaviour was characterized by the 3rd principal component (4.1% of the total variance). This effect was intensified on the second day compared to first. The 2nd (neglected) principal component comprised 4.5% of the total variance. This part of the variance in the data set could not be interpreted using the known sample classifications of our experimental design.

As expected the sequence of samples, which was revealed by PCA analysis, was clearly in agreement with the time of sampling during the first day. During the second day, the time of sampling allowed only a rough prediction of the respective kinetic metabolic phenotype (Fig. 8). Nevertheless, by visual inspection all samples appeared to exhibit a common behaviour.

We hypothesized that the apparent off-set during the intervening night period may be the source of variability. During night plants not only utilize the ^{12}C-starch generated in the previous photoperiod but may also mobilize long-term ^{13}C-carbon resources, when starch is depleted. If this competition of resources is variable in individual plants, a parameter, which is indicative of the assimilated CO_2 within the soluble metabolite pools, may be applied for correction. We tested this assumption by calculation of the mean ^{13}C-enrichment from all mass fragments. Compared to the sampling time, the mean enrichment exhibited a better fit to the sequence of samples obtained by PCA analysis (cf. graphical abstract).

Table 2
Assessment of reaction order assumptions for the ^{13}C-half-life calculation of metabolite pools after short and extended periods of $^{13}CO_2$-isotope dilution

Experiment	Organ	Observation period (h)	0. Order assumption		1. Order assumption		2. Order assumption		3. Order assumption	
			r^2	p	r^2	p	r^2	p	r^2	p
Glycine (3TMS) 174 175										
Exp. 1	Leaf	72	0.494	5.5E−04	0.614	4.4E−05	0.615	4.2E−05	0.554	1.7E−04
Exp. 1	Root	72	0.582	1.5E−03	0.691	2.3E−04	0.578	1.6E−03	0.395	1.6E−02
Exp. 2	Leaf	48	0.712	4.3E−03	0.720	3.8E−03	0.624	1.1E−02	0.539	2.4E−02
Exp. 3	Leaf	28	0.801	2.6E−11	0.780	1.1E−10	0.589	7.5E−07	0.457	4.1E−05
Exp. 4	Leaf	28	0.584	1.3E−10	0.661	1.8E−12	0.621	1.9E−11	0.523	2.7E−09
Exp. 4*	Leaf	2	0.900	1.4E−11	**0.940**	1.5E−13	**0.939**	1.7E−13	0.895	2.0E−11
Glyceric acid (3TMS) 189 191										
Exp. 1	Leaf	72	0.494	5.5E−04	0.614	4.4E−05	0.615	4.2E−05	0.554	1.7E−04
Exp. 1	Leaf	72	0.422	1.1E−03	0.356	3.4E−03	0.233	2.3E−02	0.139	8.7E−02
Exp. 1	Root	72	0.973	4.1E−05	0.906	9.4E−04	0.793	7.2E−03	0.690	2.1E−02
Exp. 2	Leaf	48	0.512	3.0E−02	0.601	1.4E−02	0.615	1.2E−02	0.593	1.5E−02
Exp. 3	Leaf	28	0.583	9.1E−07	0.586	8.2E−07	0.526	5.9E−06	0.447	5.3E−05
Exp. 4	Leaf	28	0.449	5.8E−06	0.463	3.6E−06	0.434	9.3E−06	0.370	6.5E−05
Exp. 4*	Leaf	2	**0.521**	4.8E−04	**0.507**	6.3E−04	0.478	1.0E−03	0.438	2.0E−03
Malic acid (3TMS) 245 249										
Exp. 1	Leaf	72	0.671	3.2E−06	0.589	3.1E−05	0.494	2.7E−04	0.404	1.5E−03
Exp. 1	Root	72	0.887	1.9E−07	0.841	1.4E−06	0.719	4.7E−05	0.569	8.2E−04
Exp. 2	Leaf	48	0.839	5.2E−04	0.818	8.1E−04	0.731	3.3E−03	0.595	1.5E−02
Exp. 3	Leaf	28	0.726	2.3E−09	0.635	1.4E−07	0.547	3.0E−06	0.478	2.3E−05
Exp. 4	Leaf	28	0.690	2.1E−10	0.621	7.0E−09	0.542	2.0E−07	0.458	4.3E−06
Exp. 4*	Leaf	2	**0.982**	3.1E−16	**0.981**	3.9E−16	**0.980**	5.8E−16	**0.979**	1.0E−15
Fructose (1MEOX) (5TMS) 307 310										
Exp. 1	Leaf	72	n.d.		n.d.		n.d.		n.d.	
Exp. 1	Root	72	0.715	5.3E−04	0.826	4.2E−05	0.697	7.2E−04	0.471	1.4E−02
Exp. 2	Leaf	48	0.691	5.5E−03	0.708	4.4E−03	0.530	2.6E−02	0.383	7.5E−02
Exp. 3	Leaf	28	0.673	2.9E−08	0.504	1.1E−05	0.310	1.4E−03	0.197	1.4E−02
Exp. 4	Leaf	28	0.023	3.7E−01	0.025	3.5E−01	0.026	3.4E−01	0.024	3.6E−01
Exp. 4*	Leaf	2	0.594	1.1E−04	**0.676**	1.6E−05	**0.684**	1.3E−05	0.614	7.3E−05

Linear fit of a 0, 1st, 2nd, and 3rd reaction order assumption was tested by correlation coefficient (r^2) and significance (p). Representative metabolites are shown and optimum results of short-term monitoring indicated bold (n.d., not determined). Data are from four independent experiments, Exp. 1 4. Exp. 4* represents the 2 h subset of Exp. 4.

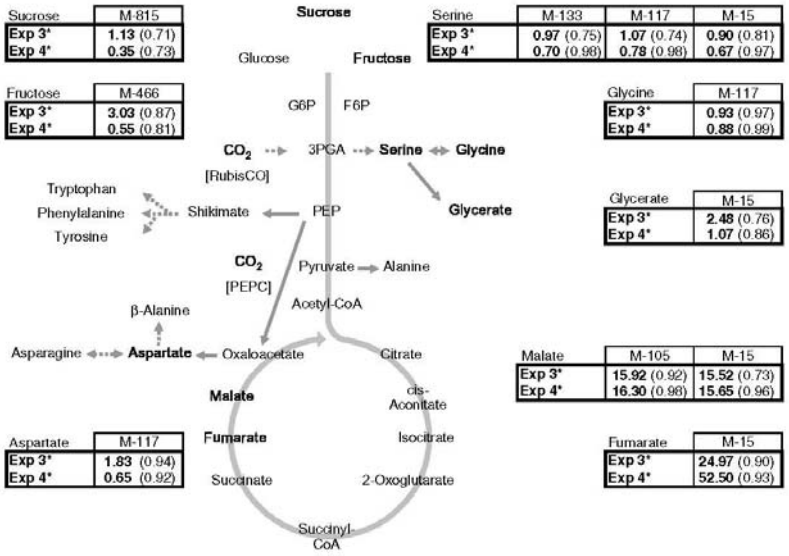


Fig. 9. Estimation of ^{13}C-half-life in metabolite pools of *Arabidopsis thaliana* shoots using a 2 h kinetic monitoring period. Experimental repeatability is indicated using the short-term results of two independent experiments, namely Exp. 3 and 4 (cf. Table 2). Exemplary results using alternative mass fragments of selected metabolite derivatives are shown. Half-life is expressed in decimal hour format and the respective regression coefficient, r^2, is reported in brackets.

Even though detailed kinetic analysis appeared to be impaired from the second photoperiod onwards we concluded, that data normalization using an indicator of total *de novo* carbon assimilation may alleviate the restriction to one photoperiod. This aspect was, however, not further investigated in the course of this study.

2.7. Estimation of the ^{13}C-half-life of metabolite pools

The development of suitable mathematical models for the presented empirical observation of ^{13}C-dilution kinetics and carbon-partitioning in higher plants is not a trivial task. As a first and strongly simplifying approach we assessed the calculation of ^{13}C-half-life of metabolite pools. Half-life calculations require assumption of reaction order. We tested 0, 1st, 2nd, and 3rd reaction order assumptions on both short-term (up to 2 h) and extended (up to 3 days) monitoring of isotope dilution after $^{13}CO_2$ labelling (Table 2). The choice of reaction order had a small effect on the quality of linear fit as determined by correlation coefficient and significance. In most cases, the 1st reaction order assumption was slightly superior and was therefore used. The analysis demonstrated that half-life calculation is indeed a simplification and may fail in some experiments or for single metabolites (cf. fructose in experiment 4; Table 2). Furthermore, the quality of fit was time course- and organ-dependent, for example glyceric acid in experiment 1 (Table 2).

Using the half-life parameter, we tested the experimental repeatability of short-term kinetic results in two independent experiments (Fig. 9). A good repeatability was found for most metabolites and ^{13}C-half-life ranging between 20 min and more than 52 h were detected. Nevertheless, we observed variability between experiments, for example sucrose and fructose (Fig. 9), which may result from slight environmental or seasonal changes. As a consequence the use of a reference genotype and highly standardized growth conditions is indicated for future comparative studies (see below).

3. Concluding remarks

We demonstrated the potential of routine GC-EI-TOF-MS profiling for flux analysis and provided an initial set of mass fragments to monitor flux and isotope dilution experiments. This set of mass fragments allows access to the steady state, but more importantly to the transient kinetics of isotope dilution at few defined carbon positions and many substructures of the targeted metabolites. Positional information can typically only be obtained by NMR studies but not through mass isotopomer analyses of molecular ions. Compared to NMR analyses the information from GC-EI-TOF-MS fragment analyses will always be limited. Only part of the carbon positions or substructures will be accessible due to the limitations of the fragmentation reac-

tions. In addition, not all mass fragment observations can be exploited, because of frequent co-elution of compounds in complex mixtures. An extension and refinement of the method can be expected, especially through improved separation of complex mixtures using GC×GC-EI-TOF-MS. The method in general will continuously benefit from the synergy with the technical advances made in GC-EI-TOF-MS metabolite profiling technology, specifically the continued substance and fragment elucidation process.

The current major limitation of our method is the lack of information on the metabolic pool sizes. The half-life of ^{13}C-label in metabolic pools is dependent on the respective pool sizes. Metabolite concentrations differ for example in response to diurnal changes. Using the sum of all mass isotopomers, which represent a fragment or molecular ion, we were able to detect relative concentration changes during the time courses we recorded in our experiments (data not shown). The estimation of exact metabolite concentrations at each time point and from each plant is a clear necessity for the improved interpretation of our primary data, as metabolite pools typically change between, species, genotypes, individuals and environmental conditions. Quantification of GC–MS analyses is routinely performed using external calibration and will be included in our future experiments. The underlying assessment of metabolite recovery, linear range and quantification using GC–MS based metabolite profiling has been reported previously (e.g. Roessner et al., 2001; Roessner-Tunali et al., 2003; Schauer et al., 2005b).

The use of populations of genetically identical, replicate plants for flux analyses was assessed. We demonstrated that the monitoring of isotope dilution after $^{13}CO_2$ labelling is feasible and that the initial variability of the plant population can be adjusted to be within the technical error of the analytic GC–MS system. However, accurate kinetic measurements are currently restricted to the initial monitoring hours. With a clear understanding of the potential and current limitations of the method mentioned above, we now head towards a phenotyping tool, which may be suitable for the estimation of carbon-partitioning into a range of representative shoot and root pools of model plant species. We suggest that the analysis of genetically modified or mutant plants may be feasible using a pair wise comparative experimental design with mixed populations of two genotypes, namely a reference genotype such as *A. thaliana* Col-0 and the genotype of interest. Both genotypes will be simultaneously grown, labelled by $^{13}CO_2$, and sampled after initiation of isotope dilution. In analogy to the concept of metabolite profiling (Fiehn et al., 2000a; Roessner et al., 2000) we envision that the half-life of ^{13}C-label in the monitored metabolic pools can be expressed as a relative change compared to the reference genotype. Thus the reference genotype will ensure comparability between independent experiments. In a similar design the adaptation of plants to physico-chemical or nutrient stresses may be feasible by simultaneous cultivation and monitoring of isotope dilution after $^{13}CO_2$ labelling of one genotype under standardized and in parallel under modified hydroponic conditions.

The presented numerical analysis of our empirical data is still in its infancy. Half-life calculations of isotope dilution after $^{13}CO_2$ labelling represent only a first, roughly simplifying approach of data reduction. We currently neglect the possibility of non-homogenous labelling and potential means to correct for the plant to plant variability. In addition the transient labelling kinetics of the different mass isotopomers, which may deviate from the expected homogenous isotope dilution behaviour, remains unexplored.

We are intrigued by our observation of quasi steady states, which do not approach the ambient input $^{12}C/^{13}C$ ratio in several subsequent photoperiods. This finding might be characteristic for whole organ analysis of higher plants and may be explained by the presence of large "inert pools" of metabolites, e.g. in the vacuole, which we currently do not separate from metabolically more active subcellular pools. In an alternative, but not mutually exclusive scenario previously assimilated carbon may compete with de novo assimilation. These current assumptions need to be experimentally analyzed prior to future modelling approaches.

4. Experimental

4.1. Chemicals and plants

The seeds of the model plant, *O. sativa* L. IR57111, were obtained from the International Rice Research Institute (IRRI; Los Baños, Philippines). *A. thaliana* L. Columbia-0 (Col-0) was a gift from Prof. T. Altmann (University of Potsdam, Germany).

Carbon dioxide (CO_2, isotopic purity 99 atom% ^{13}C, <3 atom% ^{18}O) was purchased from Sigma–Aldrich (Steinheim, Germany). All other chemicals were obtained from the same company and of highest available purity.

4.2. GC-EI-TOF-MS profiling analysis

Complete shoot material of *O. sativa* and of *A. thaliana* (rosette leaves) was sampled. Root material was taken from hydroponic culture, rinsed under tap water and soaked dry on filter paper. Metabolic inactivation was by shock freezing in liquid nitrogen. Time until frozen was 15–45 s. Metabolite extraction and chemical derivatization, namely methoxyamination and subsequent trimethylsilylation was as initially suggested by Fiehn et al. (2000a) and Roessner et al. (2000). GC-EI-TOF-MS profiling was performed using a FactorFour VF-5 ms capillary column, 30 m length, 0.25 mm inner diameter, 0.25 μm film thickness with a 10 m EZ-guard pre-column (Varian BV, Middelburg, Netherlands), and an Agilent 6890N gas chromatograph with splitless injection and electronic pressure control (Agilent, Böblingen, Germany) mounted to a Pegasus III time-of-flight

mass spectrometer (LECO Instrumente GmbH, Mönchengladbach, Germany). The initial method was slightly modified and adapted for automated GC-EI-TOF-MS analysis of 30–60 mg (FW) plant material (Wagner et al., 2003; Erban et al., 2006). Metabolites were quantified after mass spectral deconvolution (ChromaTOF software version 1.00, Pegasus driver 1.61, LECO, St. Joseph, MI, USA). The peak height representing arbitrary mass spectral ion currents of each mass isotopomer was used for subsequent numerical analysis. Mass fragment identification and retrieval of respective mass isotopomer distributions was manually supervised.

4.3. GC-EI-TOF-MS compound identification

Metabolites were identified using the NIST05 mass spectral search and comparison software (National Institute of Standards and Technology, Gaithersburg, MD, USA; http://www.nist.gov/srd/mslist.htm) and the mass spectral and retention time index (RI) collection (Schauer et al., 2005a) of the Golm Metabolome Database (GMD; Kopka et al., 2005). Only mass spectra of concurrently analyzed non-labelled plant reference material from root of the respective plant species was used for mass spectral matching. During experimentation an authenticated set of fully labelled ^{13}C-mass spectra was obtained, which was used to crosscheck identification using samples, which were fully ^{13}C-labelled. Automated mass spectral matching was manually supervised and matches accepted with thresholds set to match factor >650 (with maximum match equal to 1000) and RI deviation <1.0%. Information on the chemical derivatives and mass spectra reported in this study, may be found at http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/msri/gmd_smq.html, using the indicated mass spectral identifiers (MPIMP-ID; cf. Supplementary file 1).

4.4. Methods for the maximum $^{13}CO_2$-labelling of plants

Arabidopsis thaliana Col-0 seeds were sterilized by 10 min exposure to 70% ethanol, stratification was 2–3 days at 4 °C. Seed were germinated and pre-grown on 0.75–0.80% Noble Agar with 0.5-fold concentrated hydroponic medium (Gibeaut et al., 1997) using micro-centrifuge tubes. After one week the bottoms of the micro-centrifuge tubes were cut off and fixed onto hydroponic medium. For the following two weeks plants were grown under ambient CO_2 conditions in an 8-h day at 20 °C/18 °C (day/night) air temperature, 60–75% humidity, 120–180 μmol (photons) m^{-2} s^{-1}. At approximately the four-leaf stage, a developmentally homogenous population of up to 60 single plants was transferred with freshly prepared and degassed hydroponic solution to an atmospherically controlled, airtight 40 L growth chamber, in the following called biobox (GMS Gaswechsel-Messsysteme GmbH, Berlin, Germany). The biobox was customized for isotope specific $^{12}CO_2$ and $^{13}CO_2$ measurement and control at ambient or elevated atmospheric pressure. Monitoring options for O_2, atmospheric pressure, temperature, light intensity and relative humidity measurements were available. The set points for growth of *A. thaliana* were 10 h/14 h short day with 23 °C/18 °C air temperature, 19 °C/14 °C dew point temperature, maximum light intensity 150 μmol m^{-2} s^{-1}. The internal atmospheric pressure was dynamically kept 20 mbar above ambient. The controlled atmospheric composition was 21% oxygen and total CO_2 partial pressure set to 800 ppm. During $^{13}CO_2$-labelling a 20 ppm maximum $^{12}CO_2$-threshold was set. Above threshold a non-selective CO_2-absorption trap was used to remove emitted $^{12}CO_2$ in the light period. During night the CO_2-absorption trap was used to remove all atmospheric CO_2 from the biobox to deplete the system of $^{12}CO_2$ released by respiration (Fig. 2; cf. legend to Fig. 5 for data on the level of environmental parameter control). Isotope dilution after $^{13}CO_2$ labelling was achieved by exposure of the growth chamber to the ambient atmosphere and initial manual venting. The CO_2-isotope control settings were reversed and internal air ventilation required for the maintenance of temperature and light settings were found to be sufficient for the subsequent sampling period. The total time required for one experiment including pre-growth of a population of 60 plantlets was 45–50 days with approximately 33–35 days required for full initial labelling (e.g. Fig. 2).

Oryza sativa L. IR57111 was germinated four days in the dark at 28 °C. The germinated embryo was removed from the caryopsis and fixed to a solid lattice with the root submerged in hydroponic medium (Yang et al., 1994). Generation of developmentally homogenous populations of manipulated rice seedling required thorough preparative expertise. The growth conditions of *O. sativa* were modified to 12 h/12 h equal day and night length with constant 26 °C air temperature, 22 °C dew point temperature and maximum light intensity set to 290 μmol m^{-2} s^{-1}. All other conditions were similar to the experimental settings for *A. thaliana*. In this study two potential heterotrophic carbon sources, namely the agar and 1.0 mM (Yang et al., 1994) or 0.07 mM (Gibeaut et al., 1997) ethylenediaminetetraacetate, were neglected.

4.5. Enzyme activities

Enzyme extracts were sampled throughout four photoperiods following the start of isotope dilution after $^{13}CO_2$ labelling and prepared as described previously (Gibon et al., 2004), except 1% Triton-X100 and 20% glycerol were used. Transketolase activity was determined as described by Gibon et al. (2004). Isocitrate dehydrogenase (NAD+) was assayed by 40 min incubation of crude extract or NADH standard in freshly prepared medium comprising 5 mM $MgSO_4$, 1 mM NAD^+ (Roche), 1 mM isocitrate (Sigma–Aldrich) in 50 mM MOPS buffer, pH 7.5. The reaction mixture was stopped by an equal volume of 0.5 M NaOH. After neutralization, NADH produced by isocitrate dehydrogenase (NAD^+) was determined by using

the previously described NAD^+-based cycling protocol (Gibon et al., 2004).

Phosphoglycerate kinase was assayed by 20 min incubation of crude extract or dihydroxyacetone phosphate (Sigma–Aldrich) standard in freshly prepared medium containing 5 mM ATP (Roche), 4 mM 3-phosphoglycerate (Sigma–Aldrich), 1 U triose phosphate isomerase (Roche), 1 U NAD-glyceraldehyde-3-phosphate dehydrogenase (Roche), 2 U glycerol-3-phosphate dehydrogenase (Roche), 5 mM DTT (Sigma–Aldrich), 0.3 mM NADH (Roche), 20 mM $MgCl_2$, 50 mM KCl, 2 mM EDTA, 0.05% Triton-X100 in 100 mM Tricine/KOH, pH 8. The reaction mixture was stopped by an equal volume of 0.5 M HCl in 100 mM Tricine/KOH. After neutralization, generated glycerol-3-phosphate was determined as described (Gibon et al., 2004).

Glycerate kinase was assayed by 20 min incubation of crude extract or 3-phosphoglycerate standard in freshly prepared medium containing 10 mM $MgCl_2$, 0.05% Triton-X100, 5 mM ATP, 2 mM glycerate in 100 mM Tricine/KOH, pH 8. The reaction mixture was stopped by an equal volume 0.5 M HCl in 100 mM Tricine/KOH. After 10 min at room temperature, the solution was neutralized by 0.5 M NaOH. Then generated 3-phosphoglycerate was determined spectrophotometrically at $\lambda = 340$ nm by adding an equal volume of 4 mM $MgCl_2$, 1 U triose-phosphate isomerase (Roche), 10 U phosphoglycerate kinase (Sigma), 2 U glycerol-3-phosphate dehydrogenase (Roche), 1 U NAD-glyceraldehyde-3-phosphate dehydrogenase (Roche), 5 U glycerol-3-phosphate oxidase (Roche), 10 mM ATP (Roche), and 1 mM NADH in 100 mM Tricine/KOH, pH 8.

4.6. Calculations

Standard deviation was used throughout the manuscript to characterize experimental variability. PCA was performed as described earlier using ^{13}C-enrichment instead of the logarithm of normalized response ratios, which is typically used for metabolite profile analysis (e.g. Desbrosses et al., 2005).

^{13}C-enrichment was calculated from manually retrieved and validated mass isotopomer distributions of identified and characterized EI-induced mass fragments. Apparent ^{13}C-enrichment was corrected for the bias introduced by naturally occurring stable mass isotopes of other elements or of carbon atoms, which were introduced by chemical derivatization, using infinite dimensional matrix calculus (Wahl et al., 2004) provided by the software package of the authors. This software required a MatLab (The MathWorks, Natick, USA) programming environment. Alternatively we used a more conventional subtractive method, which corrects for natural isotope abundances (e.g. Fernandez et al., 1996). Both methods yielded highly similar correction results. The difference between both correction schemes was smaller than the contribution of the technical error, which was inherent to GC-EI-TOF-MS metabolite profiling. Data management, data transformation, basic calculations and regression analysis were performed using Microsoft Office Excel 2003 options.

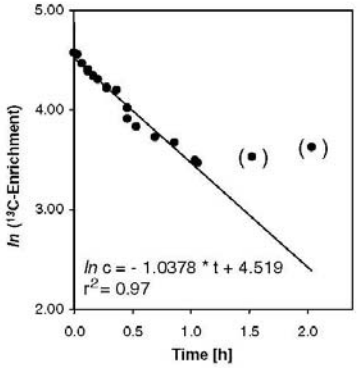


Fig. 10. Exemplary estimation of the half-life of ^{13}C-label in the serine pool using the serine (2TMS) 234 237 data shown in Fig. 7. The resulting ^{13}C-half-life is shown in Fig. 9 (serine, M-15, Exp. 4*). Empirically determined ^{13}C-enrichment was *ln*-transformed and subjected to linear regression after outlier removal. Removed measurements are indicated by brackets.

Half-life calculation was performed using the equation

$$\ln c = -k * t + \ln c_0,$$

where c represents the corrected ^{13}C-enrichment (%) and t equals time after start of isotope dilution (h). Half-life, $t_{1/2}$, was calculated from the rate constant k according to $t_{1/2} = \ln 2 * k^{-1}$. Empirical data were fitted to the above integrated rate equation by linear regression after removal of outliers (e.g. Fig. 10).

Acknowledgements

The authors acknowledge the long standing support and encouragement by Prof. L. Willmitzer and Prof. M. Stitt, Max Planck Institute of Molecular Plant Physiology, Am Muehlenberg 1, D-14476 Golm, Germany. We thank Carsten Richter (GMS Gaswechsel-Messsysteme GmbH, Berlin, Germany) for fruitful discussions and excellent technical support. This work was supported by the Max-Planck society, and the Bundesministerium für Bildung und Forschung (BMBF), Grant PTJ-BIO/0312854.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.phytochem.2007.03.026.

References

Abramson, F.P., McCaman, M.W., McCaman, R.E., 1974. Femtomole level of analysis of biogenic amines and amino acids using functional group mass spectrometry. Anal. Biochem. 57 (2), 482 499.

Baxter, C.J., Redestig, H., Schauer, N., Repsilber, D., Patil, K.R., Nielsen, J., Selbig, J., Liu, J.L., Fernie, A.R., Sweetlove, L.J., 2007. The metabolic response of heterotrophic Arabidopsis cells to oxidative stress. Plant Physiol. 143 (1), 312 325.

Bergström, K., Gürtler, J., Blomstrand, R., 1970. Trimethylsilylation of amino acids. 1. Study of glycine and lysine TMS derivatives with gas-liquid chromatography mass spectrometry. Anal. Biochem. 34 (1), 74 87.

Birkemeyer, C., Kolasa, A., Kopka, J., 2003. Comprehensive chemical derivatization for gas chromatography mass spectrometry-based multi-targeted profiling of the major phytohormones. J. Chromatogr. A 993, 89 102.

Birkemeyer, C., Luedemann, A., Wagner, C., Erban, A., Kopka, J., 2005. Metabolome analysis: the potential of in vivo labeling with stable isotopes for metabolite profiling. Trends Biotechnol. 23, 28 33.

Blaesing, O.E., Gibon, Y., Gunther, M., Hoehne, M., Morcuende, R., Osuna, D., Thimm, O., Usadel, B., Scheible, W.R., Stitt, M., 2005. Sugars and circadian regulation make major contributions to the global regulation of diurnal gene expression in Arabidopsis. Plant Cell 17 (12), 3257 3281.

Calvin, M., 1956. The photosynthetic carbon cycle. J. Chem. Soc., 1895 1915.

Calvin, M., 1964. The path of carbon in photosynthesis. In: Nobel Lectures Chemistry 1942 1962. Elsevier Publishing Company, Amsterdam, pp. 618 644.

Dauner, M., Sauer, U., 2000. GC MS analysis of amino acids rapidly provides rich information for isotopomer balancing. Biotechnol. Prog. 16, 642 649.

DeJongh, D.C., Radford, T., Hribar, J.D., Hanessia, S., Bieber, M., Dawson, G., Sweeley, C.C., 1969. Analysis of trimethylsilyl derivatives of carbohydrates by gas chromatography and mass spectrometry. J. Am. Chem. Soc. 91 (7), 1728 1740.

Desbrosses, G.G., Kopka, J., Udvardi, M.K., 2005. *Lotus japonicus* metabolic profiling. Development of gas chromatography mass spectrometry resources for the study of plant microbe interactions. Plant Physiol. 137 (4), 1302 1318.

Erban, A., Schauer, N., Fernie, A.R., Kopka, J., 2006. Non-supervised construction and application of mass spectral and retention time index libraries from time-of-flight GC MS metabolite profiles. In: Weckwerth, W. (Ed.), Metabolomics: Methods and Protocols. Humana Press, Totowa, pp. 19 38.

Fernandez, C.A., Des Rosier, C., Previs, S.F., David, F., Brunnengraber, H., 1996. Correction of ^{13}C mass isotopomer distributions for natural stable isotope abundance. J. Mass Spectrom. 31, 255 262.

Fernie, A.R., Geigenberger, P., Stitt, M., 2005. Flux an important, but neglected, component of functional genomics. Curr. Opin. Plant Sci. 8, 174 182.

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R.N., Willmitzer, L., 2000a. Metabolite profiling for plant functional genomics. Nat. Biotechnol. 18, 1157 1161.

Fiehn, O., Kopka, J., Trethewey, R.N., Willmitzer, L., 2000b. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. Anal. Chem. 72, 3573 3580.

Gibeaut, D.M., Hulett, J., Cramer, G.R., Seemann, J.R., 1997. Maximal biomass of *Arabidopsis thaliana* using a simple, low-maintenance hydroponic method and favorable environmental conditions. Plant Physiol. 115 (2), 317 319.

Gibon, Y., Blaesing, O.E., Hannemann, J., Carillo, P., Hohne, M., Hendriks, J.H.M., Palacios, N., Cross, J., Selbig, J., Stitt, M., 2004. A robot-based platform to measure multiple enzyme activities in Arabidopsis using a set of cycling assays: comparison of changes of enzyme activities and transcript levels during diurnal cycles and in prolonged darkness. Plant Cell 16 (12), 3304 3325.

Gibon, Y., Usadel, B., Blaesing, O.E., Kamlage, B., Hoehne, M., Trethewey, R., Stitt, M., 2006. Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in Arabidopsis rosettes. Genome Biol. 7 (8). Art. 76.

Kell, D.B., Brown, M., Davey, H.M., Dunn, W.B., Spasic, I., Oliver, S.G., 2005. Metabolic footprinting and systems biology: The medium is the message. Nat. Rev. Microbiol. 3, 557 565.

Klapa, M.I., Aon, J.-C., Stephanopoulos, G., 2003. Systematic quantification of complex metabolic flux networks using stable isotopes and mass spectrometry. Eur. J. Biochem. 270, 3525 3542.

Kopka, J., 2006a. Gas chromatography mass spectrometry. In: Nagata, T., Lörz, H., Widholm, J.M. (Eds.), Biotechnology in Agriculture and Forestry. In: Saito, K., Dixon, R.A., Willmitzer, L. (Eds.), Plant Metabolomics, vol. 57. Springer-Verlag, Berlin, Heidelberg, New York, pp. 3 20.

Kopka, J., 2006b. Current challenges and developments in GC MS based metabolite profiling technology. J. Biotechnol. 124, 312 322.

Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmüller, E., Dörmann, P., Gibon, Y., Stitt, M., Willmitzer, L., Fernie, A.R., Steinhauser, D., 2005. GMD@CSBDB: The Golm metabolome database. Bioinformatics 21, 1635 1638.

Laine, R.A., Sweeley, C.C., 1973. *O*-Methyl oximes of sugars analysis of *O*-trimethylsilyl derivatives by gas-liquid chromatography and mass spectrometry. Carbohydr. Res. 27 (1), 199 213.

Leimer, K.R., Rice, R.H., Gehrke, C.W., 1977. Complete mass-spectra of per-trimethylsilylated amino-acids. J. Chromatogr. 141 (3), 355 375.

Lisec, J., Schauer, N., Kopka, J., Willmitzer, L., Fernie, A.R., 2006. Gas chromatography mass spectrometry-based metabolite profiling in plants. Nat. Protocols 1, 387 396.

MacLeod, J.K., Flanigan, I.L., Williams, J.F., Collins, J.G., 2001. Mass spectrometric studies of the path of carbon in photosynthesis: positional isotopic analysis of C-13-labelled C-4 to C-7 sugar phosphates. J. Mass Spectrom. 36 (5), 500 508.

Ratcliffe, R.G., Shachar-Hill, Y., 2006. Measuring multiple fluxes through plant metabolic networks. Plant J. 45 (4), 490 511.

Roessner, U., Wagner, C., Kopka, J., Trethewey, R.N., Willmitzer, L., 2000. Simultaneous analysis of metabolites in potato tuber by gas chromatography mass spectrometry. Plant J. 23, 131 142.

Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L., Fernie, A.R., 2001. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. Plant Cell 13 (1), 11 29.

Roessner-Tunali, U., Hegemann, B., Lytovchenko, A., Carrari, F., Bruedigam, C., Granot, D., Fernie, A.R., 2003. Metabolic profiling of transgenic tomato plants overexpressing hexokinase reveals that the influence of hexose phosphorylation diminishes during fruit development. Plant Physiol. 133 (1), 84 99.

Roessner-Tunali, U., Liu, J.L., Leisse, A., Balbo, I., Perez-Melis, A., Willmitzer, L., Fernie, A.R., 2004. Kinetics of labelling of organic and amino acids in potato tubers by gas chromatography mass spectrometry following incubation in C-13 labelled isotopes. Plant J. 39 (4), 668 679.

Sanz, M.L., Sanz, J., Martinez-Castro, I., 2002. Characterization of *O*-trimethylsilyl oximes of disaccharides by gas chromatography mass spectrometry. Chromatographia 56, 617 622.

Schaefer, J., Stejskal, E.O., Beard, C.F., 1975. C-13 nuclear magnetic resonance analysis of metabolism in soybeans labelled by $^{13}CO_2$. Plant Physol. 55 (6), 1048 1053.

Schaefer, J., Kier, L.D., Stejskal, E.O., 1980. Characterization of photorespiration in intact leaves using C-13 dioxide labeling. Plant Physiol. 65 (2), 254 259.

Schauer, N., Steinhauser, D., Strelkov, S., Schomburg, D., Allison, G., Moritz, T., Lundgren, K., Roessner-Tunali, U., Forbes, M.G., Willmitzer, L., Fernie, A.R., Kopka, J., 2005a. GC MS libraries for the rapid identification of metabolites in complex biological samples. FEBS Lett. 579, 1332 1337.

Schauer, N., Zamir, D., Fernie, A.R., 2005b. Metabolic profiling of leaves and fruit of wild species tomato: a survey of the *Solanum lycopersicum* complex. J. Exp. Bot. 56 (410), 297 307.

Schwender, J., Goffman, F., Ohlrogge, J.B., Shachar-Hill, Y., 2004. Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds. Nature 432 (7018), 779 782.

Sinha, A.E., Prazen, B.J., Synovec, R.E., 2004a. Trends in chemometric analysis of comprehensive two-dimensional separations. Anal. Bioanal. Chem. 378, 1948 1951.

Sinha, A.E., Hope, J.L., Prazen, B.J., Nilsson, E.J., Jack, R.M., Synovec, R.E., 2004b. Algorithm for locating analytes of interest based on mass spectral similarity in GC×GC-TOF-MS data: analysis of metabolites in human infant urine. J. Chromatogr. A 1058, 209 215.

van Winden, W.A., van Dam, J.C., Ras, C., Kleijn, R.J., Vinke, J.L., van Gulik, W.M., Heijnen, J.J., 2005. Metabolic-flux analysis of *Saccharomyces cerevisiae* CEN.PK113-7D based on mass isotopomer measurements of 13C-labeled primary metabolites. FEMS Yeast Res. 5, 559 568.

Wagner, C., Sefkow, M., Kopka, J., 2003. Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. Phytochemistry 62 (6), 887 900.

Wahl, S.A., Dauner, M., Wiechert, W., 2004. New tools for mass isotopomer data evaluation in ^{13}C flux analysis: mass isotope correction, data consistency checking and precursor relationships. Biotechnol. Bioeng. 85 (3), 259 268.

Wittmann, C., Heinzle, E., 2001. Application of MALDI-TOF MS to lysine-producing *Corynebacterium glutamicum*. A novel approach for metabolic flux analysis. Eur. J. Biochem. 268, 2441 2455.

Yang, X., Romheld, V., Marschner, H., 1994. Effect of bicarbonate on root-growth and accumulation of organic-acids in Zn-inefficient and Zn-efficient rice cultivars (*Oryza sativa* L.). Plant Soil 164, 1 7.

Appendix B: Metabolomic Software and Database Development

[8] Wagner C, Sefkow M, **Kopka J** (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry* 62(6): 887-900 (http://dx.doi.org/10.1016/S0031-9422(02)00703-3) (http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6TH7-47WBXD4-B&_user=10&_coverDate=03%2F31%2F2003&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=8b9063bba35ffa0c37784e717f43167e)

Project leadership and concept development towards establishment of a combined mass spectral and retention index reference library. The requirement of reference libraries for metabolite identification in complex GC-MS profiles was demonstrated and respective methods and technologies for the acquisition of relevant physicochemical properties, later termed mass spectral tags (MSTs), were reported. A detailed description of a standardized procedure was published later as a book chapter (cf. above, **Erban et al. 2007 [4]**).

[9] Luedemann A, Weicht D, Selbig J, **Kopka J** (2004) PaVESy: pathway visualization and editing system. *Bioinformatics* 20 (16): 2841-2844 (http://dx.doi.org/10.1093/bioinformatics/bth278) (http://bioinformatics.oxfordjournals.org/cgi/content/abstract/20/16/2841) Free article (http://bioinformatics.oxfordjournals.org/cgi/reprint/20/16/2841)

Project leadership of the software design for the interactive visualization and editing of metabolic pathway data. The PaVESy tool was created to support the interpretation of metabolite profiling results in the context of metabolic networks.

[10] Schauer N, Steinhauser D, Strelkov S, Schomburg D, Allison G, Moritz T, Lundgren K, Roessner-Tunali U, Forbes MG, Willmitzer L, Fernie AR, **Kopka J** (2005) GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Letters* 579 (6): 1332-1337 (http://dx.doi.org/10.1016/j.febslet.2005.01.029) (http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6T36-4FBW3PM-3&_user=10&_coverDate=02%2F28%2F2005&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=8b296eb216120ee8f3dc73c04120628c)

[11] **Kopka J**, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmueller E, Doermann P, Weckwerth W, Gibon Y, Stitt M, Willmitzer L, Fernie AR, Steinhauser D (2005) GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* 21 (8): 1635-1638 (http://dx.doi.org/10.1093/bioinformatics/bti236) (http://bioinformatics.oxfordjournals.org/cgi/content/abstract/21/8/1635) Free article (http://bioinformatics.oxfordjournals.org/cgi/content/full/21/8/1635)

Project leadership of an international collection of mass spectral and retention index reference data of methoxyaminated and silylated chemical derivatives relevant for GC-MS based metabolite profiles of plant and microbial sources. This library was made publicly available through the Golm Metabolome Database (GMD), as integral part of the PhD thesis of Dr. Dirk Steinhauser. A detailed description of the GMD collection was subsequently published as a book chapter: Hummel J, Selbig J, Walther D, **Kopka J** (2008) The Golm metabolome database: a database for GC-MS based metabolite profiling. In: Nielsen J, Jewett M (eds) Metabolomics a powerful tool in systems biology. Topics in Current Genetics Vol. 18, Springer-Verlag, Berlin Heidelberg New York, pp 75-96. The GMD project is now continued in cooperation with Prof. Dr. Joachim Selbig, (University of Potsdam) and Dr. Dirk Walther (Max-Planck-Institute of Molecular Plant Physiology).

[12] Scholz M, Kaplan F, Guy CL, **Kopka J**, Selbig J (2005) Non-linear PCA: a missing data approach. *Bioinformatics* 21 (20): 3887-3895
(http://dx.doi.org/10.1093/bioinformatics/bti634)
(http://bioinformatics.oxfordjournals.org/cgi/content/abstract/21/20/3887)
Free article (http://bioinformatics.oxfordjournals.org/cgi/content/full/21/20/3887)

Co-supervision of the exploration of non-linear statistical methods aimed at improved retrieval of relevant metabolites from time course data. The motivation to this study and one of the first metabolite profiling time series data sets was contributed. The central observation of non-linear metabolite correlations in time series analyses was made and the demand for suitable missing data substitution procedures expressed. The results of these investigations were the incentive to extend the MetaGenalyse internet application by non-linear algorithms, such as the probabilistic PCA or the Bayesian PCA.

[13] Luedemann A, Strassburg K, Erban A, **Kopka J** (2008) TagFinder for the quantitative analysis of gas chromatography - mass spectrometry (GC-MS) based metabolite profiling experiments. *Bioinformatics* 24 (5): 732 -737
(http://dx.doi.org/10.1093/bioinformatics/btn023)
(http://bioinformatics.oxfordjournals.org/cgi/content/abstract/24/5/732)
Free article (http://bioinformatics.oxfordjournals.org/cgi/content/full/24/5/732)

Project leadership of the establishment of publicly available chromatography data processing software for standardized non-targeted fingerprinting analysis and metabolite targeted profiling analyses of GC-MS studies. The TagFinder software, developed by Alexander Luedemann, supports a standardized chromatography data processing workflow prerequisite for the establishment of a future database of metabolite profiles. In addition, the retrieval of mass isotopomer ratios and distributions for stable isotope tracing and flux analyses is supported. TagFinder utilizes the GMD mass spectral and retention index library for compound identification.

[14] Strehmel N, Hummel J, Erban A, Strassburg K, **Kopka J** (2008) Estimation of retention index thresholds for compound matching using routine gas chromatography-mass spectrometry based metabolite profiling experiments. *J Chromatogr B* 871: 182-190
(http://dx.doi.org/10.1016/j.jchromb.2008.04.042)
(http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6X0P-4SFXK82-1&_user=10&_coverDate=08%2F15%2F2008&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=43339c6333ca556efc7f8b339723c2d4)

Project leadership of a study exploring the application of retention index information for compound identification in GC-MS based metabolite profiles. The transfer and prediction of retention index libraries between typical GC-MS chromatography systems widely applied for metabolite profiling was investigated and matching thresholds deduced for the enhanced future use of the GMD compendium for compound recognition and structural elucidation of yet non-identified MSTs.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
PERGAMON
Phytochemistry 62 (2003) 887–900
PHYTOCHEMISTRY
www.elsevier.com/locate/phytochem

Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles

Cornelia Wagner[a], Michael Sefkow[b], Joachim Kopka[a,*]

[a]*Department Willmitzer, Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, D-14467 Golm, Germany*
[b]*Division of Natural Product Chemistry, Institute for Chemistry, University of Potsdam, Karl-Liebknecht-Strasse 24-25, D-14467 Golm, Germany*

Received 9 September 2002; received in revised form 18 October 2002

Abstract

The non-supervised construction of a mass spectral and retention time index data base (MS/RI library) from a set of plant metabolic profiles covering major organs of potato (*Solanum tuberosum*), tobacco (*Nicotiana tabaccum*), and *Arabidopsis thaliana*, was demonstrated. Typically 300–500 mass spectral components with a signal to noise ratio $\geqslant 75$ were obtained from GC/EI-time-of-flight (TOF)-MS metabolite profiles of methoxyaminated and trimethylsilylated extracts. Profiles from non-sample controls contained approximately 100 mass spectral components. A MS/RI library of 6205 mass spectral components was accumulated and applied to automated identification of the model compounds galactonic acid, a primary metabolite, and 3-caffeoylquinic acid, a secondary metabolite. Neither MS nor RI alone were sufficient for unequivocal identification of unknown mass spectral components. However library searches with single bait mass spectra of the respective reference substance allowed clear identification by mass spectral match and RI window. Moreover, the hit lists of mass spectral searches were demonstrated to comprise candidate components of highly similar chemical nature. The search for the model compound galactonic acid allowed identification of gluconic and gulonic acid among the top scoring mass spectral components. Equally successful was the exemplary search for 3-caffeoylquinic acid, which led to the identification of quinic acid and of the positional isomers, 4-caffeoylquinic acid, 5-caffeoylquinic acid among other still non-identified conjugates of caffeic and quinic acid. All identifications were verified by co-analysis of reference substances. Finally we applied hierarchical clustering to a complete set of pair-wise mass spectral comparisons of unknown components and reference substances with known chemical structure. We demonstrated that the resulting clustering tree depicted the chemical nature of the reference substances and that most of the nearest neighbours represented either identical components, as judged by co-elution, or conformational isomers exhibiting differential retention behaviour. Unknown components could be classified automatically by grouping with the respective branches and sub-branches of the clustering tree.

Keywords: *Arabidopsis thaliana*; Brassicaceae; *Solanum tuberosum*; *Nicotiana tabacum*; Solanaceae; Caffeoylquinic acids; Chlorogenic acid; Caffeic acid; Quinic acid; Galactonic acid; Gluconic acid; Glucaric acid; Gulonic acid; Ascorbic acid; Metabolite profiling; Mass spectral library; Gas chromatography (GC); Time-of-flight mass spectrometry (TOF-MS); Retention time index (RI)

1. Introduction

In the next few years metabolome analyses will have emerged from infancy to being firmly established as the third cornerstone of functional genomics. Multi-parallel measurements of the large variety of primary and secondary metabolites ideally complement the current focus of functional genomics, namely mRNA profiling and proteomics approaches. Metabolite profiling technology will spread not only within the area of plant biotechnology (Fiehn et al., 2000a; Fiehn, 2002; Frenzel et al., 2002; Huhman and Sumner, 2002; Roessner et al., 2000, 2001a) but across all fields of biological science, because it allows direct access to general systems analyses as is exemplified by applications in drug discovery (Boros et al., 2002) or the profiling of genetic disorders (Griffîn et al., 2001).

Abbreviations: EI, Electron impact ionization; GC, gas chromatography; RI, retention time index; MS, mass spectrum; TOF, time-of-flight

* Corresponding author. Tel.: +49-331-567-8262; fax +49-331-567-898262.
E-mail address: kopka@mpimp-golm.mpg.de (J. Kopka).

0031-9422/03/$ - see front matter © 2003 Published by Elsevier Science Ltd.
doi:10.1016/S0031-9422(02)00703-3

One of the central technology platforms of metabolic profiling technology is bench-top gas chromatography coupled to mass spectrometry. The choice of this hyphenated technology was motivated by unsurpassed combination of chromatographic separation power, selectivity, sensitivity, and dynamic range of mass detection. Moreover, both gas chromatography and electron impact ionization mass spectrometry exhibit extremely high reproducibility and are now applied to multi-parallel analyses of hundreds of biological samples (Fiehn et al., 2000a). Prior and recent applications in medical diagnostics (Duez et al., 1996; Matsumoto and Kuhara, 1996; Kuhara, 2001) as well as upcoming publications within plant science also rely on GC/MS technology.

The major application of GC/MS based metabolite profiling has been phenotypic characterization and classification of genetically altered plant samples with attempts to copy the phenotype of genetic lesions by experimental treatment of non-modified plants (Roessner et al., 2001a, b). However, approaches to utilize the wealth of mass spectral information within each single GC/MS profile have not yet been explored. The final aim of this aspect is a comprehensive and automated analysis of all mass spectral components from a metabolic profile. The importance of identification and chemical classification of the mass spectral components may best be acknowledged by analogy with functional annotation of unknown genes in transcriptome analysis. Whereas genes are characterized by nucleotide sequences, electron impact ionization mass spectra (EI-MS) have a similar import for small compounds. They represent the fingerprint of the molecular fragmentation pattern of chemical structures. Annotation and classification of genes by alignment and evaluation of sequence homology today is fully automated. In parallel the comparison of mass spectra is automated to a similar degree. The NIST (National Institute of Standards and Technology, Gaithersburg, MD, USA) released a mass spectral search program which is publicly available and represents a platform independent GC/MS analysis software. However, the mass spectral match is insufficient for unequivocal substance identification, mainly because structural isomers especially conformational isomers may produce highly similar mass spectra.

Thus a typical attempt to identify a given component from a GC/MS profile is tedious. The process starts with manual interpretation of the fragmentation pattern and is supported by mass spectral comparison with commercial libraries. Pure substance of reference compounds is subsequently required to establish co-chromatography and mass spectral identity with unknown candidate components. Unfortunately the direct route via preparation of pure fractions from GC/MS runs and subsequent unambiguous identification by NMR is currently not feasible because of the enormous difference in preparative capacity and analytical sensitivity. Thus component identification from GC/MS profiles is restricted by commercial or preparative access to pure reference compounds.

As a consequence we started to develop an efficient use of those reference compounds that are available. We currently focus on the simultaneous identification and classification of unknown components in a high number of different plant matrices. For this purpose we exploit the two characteristic substance properties provided by GC/MS, i.e., retention time index (RI) and mass spectrum. Other than stand-alone mass spectrometry hyphenated techniques like GC/MS provide the potential to separate conformational isomers prior to mass detection. Thus with the aim to group and annotate unknown compounds from multiple matrices it appears to be highly negligent not to utilize the chromatographic property of natural products.

Novel software developments promise automated calculation of RI and correct extraction in other words deconvolution of mass spectra from GC/MS chromatograms. In addition recently developed GC/EI-TOF-MS technology apparently exhibited high mass spectral reproducibility (Veriotti and Sacks, 2000). With these incentives we explored the potential of the combination of both technologies for qualitative analysis of GC/MS based metabolic profiles. We describe the non-supervised construction of a mass spectral and retention time index database (MS/RI library) from routine GC/EI-TOF-MS metabolite profiles of a trial selection of typical plant matrices. The composition of this mass spectral compendium is characterized and its application in substance identification and automated classification of unknown components is discussed.

2. Results and discussion

2.1. Choice of technology

GC/EI-TOF-MS technology was chosen for the purpose of this project. Other than scanning type mass spectrometers like quadrupole, ion trap, or sector instruments, time-of-flight technology combined fast acquisition rates with mass spectral integrity. Whereas in typical scanning GC/MS datasets the relative fragment abundance of mass spectra shifted from peak front to tail, mass spectra stayed unchanged when monitored by TOF-MS. This property and the option of high-resolution data aquisition were shown to be technologically beneficial with respect to deconvolution of co-eluting components and mass spectral comparison of petrochemical (Veriotti and Sacks, 2000) as well as biological profiles (Veriotti and Sacks, 2001).

Mass spectral deconvolution and automated calculation of RI was performed by the automated mass spectral

deconvolution and identification system, AMDIS (National Institute of Standards and Technology, Gaithersburg, MD, USA). AMDIS was the first software tool available, which combined automated mass spectral deconvolution and calculation of RI. Moreover, AMDIS is publicly available and compatible with almost all file formats generated by commercial GC/MS systems including widely distributed quadrupole GC/MS systems. However, recent versions of the ChromaTOF™ software (LECO, St. Joseph, MI, USA) include fully equivalent options, but this software is restricted to the GC/EI-TOF-MS file format of the company. The choice of AMDIS in the present paper neither reflects a quality assessment nor is a software comparison intended. Alternative test versions of MS/RI libraries were successfully generated by ChromaTOF™ software version 1.6 and are in use.

AMDIS exhibited mainly two types of errors, i.e., deconvolution of more than one component per peak (Tables 3 and 4) and generation of accidental components with low intensity. Both errors were influenced by scan to scan fluctuations of the base line. Setting the smoothing option to a width of four scans and removal of components with a signal to noise ratio <75 empirically reduced both errors. A statistical assessment of residual erroneous components and of the number of true components lost was impossible, because all attempts to automatically identify artefact mass spectra by common properties failed. The tested parameter comprised the mass spectral composition and all available validation parameter of peak deconvolution which were provided by the AMDIS software. The NIST98 mass spectral search program (National Institute of Standards and Technology, Gaithersburg, MD, USA) is fully integrated with both the AMDIS and the ChromaTOF™ deconvolution software and, therefore, was chosen as software platform for MS/RI libraries.

The metabolite profiles entering the present MS/RI library were from a single series of GC/MS runs. Profiles from multiple analytical series may be combined (data not shown), however, the possibility of slight accidental changes of RI or mass spectral tuning settings was avoided in the present work. Gas chromatography was optimised to cover the volatility range of *n*-dodecane to *n*-hexatriacontane within a temperature ramp of 15 min. The data acquisition rate was set to 6 spectra per second resulting in 15 20 data points across a peak. This setting confirmed with the limit recommended for accurate description of chromatographic peaks (Dyson, 1999; van Deursen et al., 2000). Split injection used in previous publications was replaced by a hot splitless mode of injection. This development resulted in approximately 10-fold reduction of required sample amount (Table 2).

2.2. RI/MS Library construction and composition

The present MS/RI library was composed of mass spectral components obtained from 21 metabolic profiling experiments. Sample information and resulting number of components were listed in Table 1. The MS/RI library held 6205 mass spectra exhibiting an average peak purity of 47.4% (Table 2). Of these mass spectra 5.6% were derived from control experiments without plant sample. These non-sample controls were interspersed within plant samples with the aim to check for chemical contamination and in order to monitor memory effects in the course of the GC/MS sequences. Only caffeoylquinic acids exhibited slight carry over into the final control experiment (data not shown). The 18 plant profiles were generated in equal numbers from either 1 3 mg or 10 18 mg fresh weight. The nine samples with increased fresh weight had less than twice the yield of components and exhibited a significant decrease in peak purity. Pre-experiments with still higher sample load showed not only progressively reduced peak purity but also resulted in artefact RI. RI shifts were caused by peak overloading (data not shown).

The plant samples were from potato (*Solanum tuberosum*), tobacco (*Nicotiana tabacum*) and *Arabidopsis thaliana*. They comprised 53.6, 21.9, and 18.9% of library components, respectively. Leaf and root organs predominated. For potato plants three further organs were included flower, stolon, and tuber. Root and tuber samples appeared to contain less components per fresh weight than leaf. In the case of tuber samples this finding confirms earlier observations (Roessner et al., 2000, 2001a).

A comprehensive representation of the MS/RI library with regard to the distribution of component amount and RI is given in Fig. 1. This plot may be interpreted as the full chromatographic reconstruction of all mass spectral components extracted from the present 21 metabolic profiles, where RI represents the time axis and the amount value replaces signal intensity. The amount values covered a range of four orders of magnitude, from 1.05 10^7 to 1.26 10^3, at threshold signal to noise ratio. The average signal to noise ratio was 397 accompanied by an average amount of 3.59 10^5. Coelution was apparent (Fig. 1) to such an extent that peak identification could not rely exclusively on chromatographic retention. However, RI showed high reproducibility with typical standard deviations ranging from 0.1 to 2.0 (Tables 3 and 4). Therefore we concluded that under our experimental conditions RI is a highly valid additional parameter for substance identification.

2.3. Screening for model compounds

Purification strategies applied prior to quantitative measurements typically aim at reduction of sample complexity. Clean-up procedures use selective techniques of

Table 1
Sample description of 21 metabolic profiles generated by GC/EI-TOF-MS

Sample information				Components (S/N ≥ 75)
Experiment	Species	Organ	Fresh weight (mg)	Number
1135ec05	*Arabidopsis thaliana* C24	Leaf	1	282
1135ec06	*Arabidopsis thaliana* C24	Root	2	181
1135ec07	*Nicotiana tabacum* cv. SNN	Leaf	3	295
1135ec08	*Nicotiana tabacum* cv. SNN	Root	3	226
1135ec09	*Solanum tuberosum* cv. Désirée	Leaf	1	300
1135ec10	*Solanum tuberosum* cv. Désirée	Root	3	243
1135ec13	*Solanum tuberosum* cv. Désirée	Flower	3	244
1135ec12	*Solanum tuberosum* cv. Désirée	Stolon	3	221
1135ec11	*Solanum tuberosum* cv. Désirée	Tuber	1	175
				Σ – 2167
1135ec24	*Arabidopsis thaliana* C24	Leaf	16	383
1135ec25	*Arabidopsis thaliana* C24	Root	10	327
1135ec26	*Nicotiana tabacum* cv. SNN	Leaf	11	374
1135ec27	*Nicotiana tabacum* cv. SNN	Root	10	463
1135ec28	*Solanum tuberosum* cv. Désirée	Leaf	14	503
1135ec29	*Solanum tuberosum* cv. Désirée	Root	14	366
1135ec32	*Solanum tuberosum* cv. Désirée	Flower	10	440
1135ec31	*Solanum tuberosum* cv. Désirée	Stolon	14	440
1135ec30	*Solanum tuberosum* cv. Désirée	Tuber	18	395
				Σ – 3691
1135ec03	Non-sample control [a]	Initial	Empty	119
1135ec23	Non-sample control [a]	Interspersed	Empty	76
1135ec61	Non-sample control [a]	Final	Empty	152
				Σ – 347

Mass spectral components were deconvoluted by AMDIS software. Components with signal to noise ratio < 75 were removed from further analysis.
[a] Non-sample control experiments represent empty containers for sampling and storage which were fully processed in parallel with regular samples from extraction to final GC/MS analysis.

Table 2
Global description of the mass spectral and retention time index database. Sample classes are characterized by number of components (Sum), composition (%), average number (AVG), and standard deviation (SD) of components per experiment

Sample class	Experiments	Components				Signal to noise ratio	Peak purity
		Sum	%	AVG[a]	S.D.[b]	AVG[a]	AVG[a]
Total	21	6205	100.0	295	120	397	47.4
Non-sample control	3	347	5.6	116	38	267	59.7
1–3 mg (FW)	9	2167	34.9	241	46	390	51.2
10–18 mg (FW)	9	3691	59.5	410	55	447	39.4
Solanum tuberosum cv. Désirée	10	3327	53.6	333	–	–	–
Arabidopsis thaliana C24	4	1173	18.9	293	–	–	–
Nicotiana tabacum cv. SNN	4	1358	21.9	340	–	–	–
Leaf[c]	6	2137	34.4	356	–	–	–
Root[d]	6	1806	29.1	301	–	–	–
Flower[e]	2	684	11.0	342	–	–	–
Stolon[f]	2	661	10.7	331	–	–	–
Tuber[g]	2	570	9.2	285	–	–	–

The general quality is assessed by average signal to noise ratio and peak purity. Experiments constituting the classes, non-sample control, 1–3 mg (FW), 10–18 mg (FW), and the classes of plant species are indicated in Table 1.
[a] Average.
[b] Standard deviation.
[c] Experiments 1135ec05, 07, 09, 24, 26, and 28.
[d] Experiments 1135ec06, 08, 10, 25, 27, and 29.
[e] Experiments 1135ec13 and 32.
[f] Experiments 1135ec12 and 31.
[g] Experiments 1135ec11 and 30.

Table 3
Summary of the mass spectral search list obtained with the bait mass spectrum of persilylated 3-caffeoyl quinic acid

Components Identification	Number of replicates	Retention time index		Match	
		AVG[a]	S.D.[b]	AVG[a]	S.D.[b]
3-Caffeoyl quinic acid (BP)	9	2990.3	0.7	891	96
3-Caffeoyl quinic acid[d]	24[c]	3113.7	2.0	843	132
4-Caffeoyl quinic acid (BP)	7	3009.1	1.2	783	34
5-Caffeoyl quinic acid[d]	15	3190.0	0.8	772	99
4-Caffeoyl quinic acid[d]	20	3168.3	0.9	738	84
Quinic acid	23[c]	1854.9	1.0	701	72
A1	4	3096.8	0.5	671	88
A2	4	2982.8	0.4	667	108
A3	5	2851.8	0.4	644	96
A4	4	2924.5	0.3	600	70
A5	3	2936.0	0.4	589	13
A6	17	1715.2	0.7	574	32
Caffeic acid	13	2139.0	0.4	560	24
A7	20	1809.1	1.3	559	28
A8	3	2856.2	1.5	552	57
p-Coumaric acid[d]	12	1947.0	0.6	550	27
A9	4	2587.9	0.3	543	28
A10	10	1756.4	0.4	542	5
Galactonic acid	9	1991.7	0.7	533	8
A11	9	1766.5	0.4	532	20
1-Caffeoyl quinic acid	1	3397.0	–	710	–
4-Caffeoyl quinic acid (BP)[d]	2	3138.7	–	624	–
1-Caffeoyl quinic acid (BP)	1	3296.6	–	619	–
1-Caffeoyl quinic acid (BP)	1	3152.5	–	586	–
Caffeic acid (BP)	1	1985.5	–	564	–
5-Caffeoyl quinic acid (BP)	1	3007.7	–	527	–
Ascorbic acid	1	1946.4	–	527	–
Isoascorbic acid	1	1957.4	–	521	–
o-Coumaric acid	1	1821.6	–	497	–

Twenty groups of components are shown (group size ⩽3, match factor ⩽532). Identifications were confirmed manually. Tested reference substances which were not observed in plant samples are appended (BP; by-product of chemical synthesis or derivatization).
[a] Average.
[b] Standard deviation.
[c] Deconvolution of more than one component per peak.

Table 4
Summary of the mass spectral search list obtained with the bait mass spectrum of persilylated galactonic acid

Components Identification	Number of replicates	Retention time index		Match	
		AVG[a]	S.D.[b]	AVG[a]	S.D.[b]
Galactonic acid	19	1991.6	0.7	893	79
Gluconic acid	11	1997.0	1.1	849	54
B1	17	2010.9	0.5	836	54
B2	12	1986.6	0.8	820	45
Gulonic acid[d]	8	1959.2	0.7	786	52
B3	8	1766.4	0.5	783	54
B4	4	1700.2	0.1	766	11
Ribitol[e]	21	1727.1	0.3	762	12
Mannitol	15	1927.7	0.6	761	34
B5	5	1744.8	0.2	760	33
B6	11	1919.7	1.0	754	30
Threonic acid	24[c]	1559.6	0.7	754	25
B7	11	1756.3	0.4	750	41
Glucose methoxyamine (BP)	16	1908.5	0.8	746	17
B8	8	1603.1	0.4	740	21
B9	5	1972.5	0.4	738	18
B10	8	1541.0	0.4	736	32
Galactose methoxyamine	6	1883.7	0.7	734	13
B11	13	2766.1	1.3	732	27
B12	17	2105.0	0.8	731	25
B13	11	2128.0	0.3	726	25
B14	3	2290.9	0.2	722	41
Glucose methoxyamine	17	1889.8	1.4	721	21
B15	9	2041.3	0.3	720	27
Sorbitol	1	1931.1	–	763	–
Galactitol	1	1935.5	–	753	–
Glucaric acid	1	2014.0	–	743	–
Xylitol	1	1710.0	–	732	–
Ascorbic acid	1	1946.4	–	632	–
Isoascorbic acid	1	1957.4	–	637	–

Twenty four groups of components are shown (group size ⩽3; match factor ⩽720). Identifications were confirmed manually. Tested reference substances which were not observed in plant samples are appended (BP; by-product of chemical synthesis or derivatization).
[a] Average.
[b] Standard deviation.
[c] Deconvolution of more than one component per peak.
[d] Identified as by-product of D-gulonic acid gamma-lactone.
[e] Internal standard.

extraction and enrichment. Finally a small number of target compounds which exhibit unequivocal chromatographic separation are prepared and analyzed. In contrast metabolic profiling utilizes substance specific and in the context of complex samples selective means of detection. For example selective ions after mass fragmentation (Fiehn et al., 2000a; Huhman and Sumner, 2002) or characteristic chemical shifts within NMR spectra (Fan et al., 2001) are exploited in order to resolve complex chromatograms into clear substance identifications. An example of the chromatographic separation obtained by GC/EI-TOF-MS is shown in Fig. 2. In this potato leaf matrix the persilylated 3-caffeoylquinic acid derivative was accompanied by 3 co-eluting components exhibiting the unique masses $m/z = 259$ and $m/z = 597$. Different to scanning GC/MS chromatograms abundant and specific TOF mass traces of the 3-caffeoylquinic acid derivative, $m/z = 255$, 307, and 345 (Fuchs and Spiteller, 1996), exhibited complete overlap when scaled to maximum. Even an ion like $m/z = 447$, which in the case of the 3-caffeoylquinic acid derivative had a mass spectral proportion of less than 0.1%, showed deviation only of shape but not of maximum intensity of the chromatographic trace. All mass spectra mentioned in the following are implicitly of substances which were subject to methoxyamination and trimethylsilylation.

In contrast to chemical pre-purification or selected ion processing of complex GC/EI-TOF-MS profiles we introduced selectivity to our analyses by generating subsets of MS/RI library components by exploiting mass spectral similarity with single bait mass spectra. Hit lists of mass spectral searches are routinely ordered according to mass spectral match factors. Here we present the 400

highest ranking results of mass spectral searches in analogy to Fig. 1 as two dimensional plots with one axis representing RI and the second axis used to describe the respective match factor of the library component (Figs. 3 and 4).

2.3.1. Exemplary search for 3-caffeoylquinic acid

The top ranking mass spectral components extracted from the MS/RI library by comparison with the mass spectrum of 3-caffeoylquinic acid ranged from match factor 510 to the best fit of 949. Almost the complete

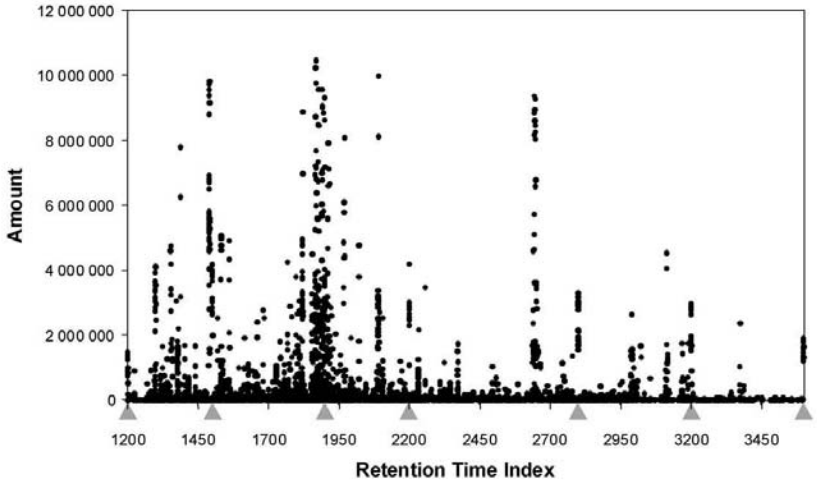


Fig. 1. Full chromatographic reconstruction of all mass spectral components comprising the mass spectral and retention time index database. Arrows indicate the chromatographic position of the retention time standards, *n*-dodecane (RI 1200), *n*-pentadecane (RI 1500), *n*-nonadecane (RI 1900), *n*-docosane (RI 2200), *n*-octacosane (RI 2800), *n*-dotriacontane (RI 3200), and *n*-hexatriacontane (RI 3600).

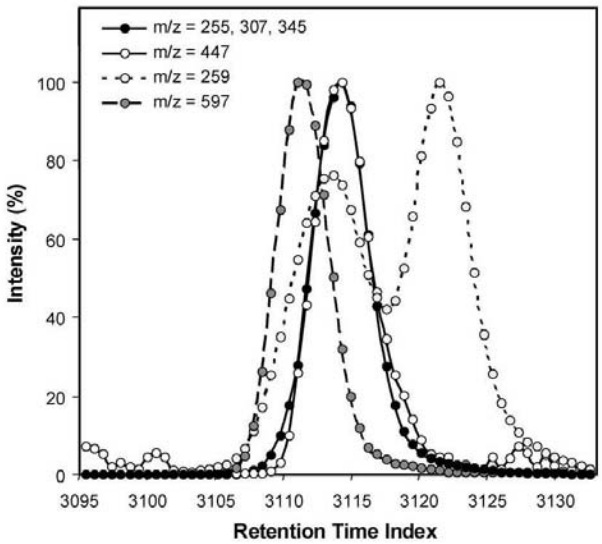


Fig. 2. Partial selected ion chromatogram of persilylated 3-caffeoylquinic acid (m/z = 255, 307, 345, 447) and co-eluting components (m/z = 597 and 259) from potato leaf. Ion intensity was scaled to maximum within the selected RI window. Different to scanning GC/MS chromatograms the single TOF mass traces of m/z = 255, 307, and 345 exhibited complete overlap. The trace m/z = 447 exemplifies a specific fragment with a mass spectral proportion <0.1%. The figure shows GC/EI-TOF-MS experiment 1135ec28 (refer to Table 1).

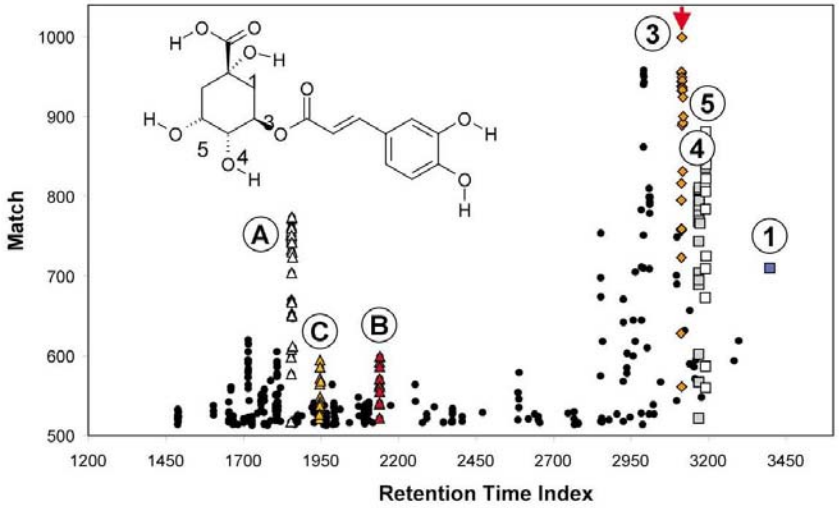


Fig. 3. Chromatographic reconstruction of the mass spectral search list of plant components and reference substances obtained with the bait mass spectrum of persilylated 3-caffeoylquinic acid (arrow). The search list contained mass spectra of the sialated derivatives of the monomers as well as positional isomers of the 3-caffeoylquinic acid conjugate. A (open triangle) quinic acid, B (red triangle) caffeic acid, C (orange triangle) *p*-coumaric acid, 1 (blue box) 1-caffeoylquinic acid, 3 (orange diamond) 3-caffeoylquinic acid, 4 (grey box) 4-caffeoylquinic acid, and 5 (open box) 5-caffeoylquinic acid. Replicate mass spectra from plant samples were confirmed manually. Silylated protons and ester bond position of the respective isomers are indicated on the inset chemical structure.

range of RI was covered. Twenty-three mass spectral components from plant samples were detected and manually confirmed to represent 3-caffeoylquinic acid. This compound was found in all samples of the solanaceous species and in root samples of *Arabidopsis thaliana*. The average match factor of this group of mass spectral components was 843 ± 132 with RI = 3113.7 ± 2.0 (Table 3). Redundant deconvolution of single chromatographic peaks was found to be insufficiently repressed. However, this error appeared to be irrelevant or even beneficial for the positive identification of plant components.

The complete hit list was dissected into groups of components which were separated by RI gaps > 2.0. Thirty-nine of these groups had more than two members. The components of these groups were compared with reference profiles of commercially available quinic acid, caffeic acid, *o*-coumaric acid, *p*-coumaric acid, 3-caffeoylquinic acid, and synthetic preparations of the three isomers 1-caffeoylquinic acid, 4-caffeoylquinic acid, and 5-caffeoylquinic acid. Main derivatives (Fig. 3) and by-products were taken into account. Eight of the top ranking groups of components were identified to represent substances which were chemically related to the initial bait, 3-caffeoylquinic acid. In detail quinic acid, caffeic acid, *p*-coumaric acid, 3-caffeoyl-, 4-caffeoyl-, and 5-caffeoylquinic acid were found in addition to respective by-products of both commercial or synthetic preparations (Table 3). Both 4-caffeoylquinic acid and 5-caffeoylquinic acid were expected to be present in potato (Griffiths and Bain, 1997; Percival and Baird, 2000) and tobacco tissue (Koeppe et al., 1969; Baumert et al., 2001) next to the main isomer 3-caffeoylquinic acid. The deconvoluted GC/EI-TOF mass spectra of the silylated main products of 3-caffeoyl-, 4-caffeoyl-, and 5-caffeoylquinic acid were confirmed by published data including the proportions of the key ions $m/z = 307$, 345, 447 (Fuchs and Spiteller, 1996). The TOF mass detector, however, generated characteristic mass spectra with increased proportions of ions with low m/z and reduced proportions of high m/z (Supplementary file 2) as compared with mass spectra generated by high resolution sector field instruments (Fuchs and Spiteller, 1996).

Galactonic acid, ascorbic acid, and isoascorbic acid exhibited mass spectral match factors = 521 532. We could therefore empirically deduce a threshold of 500 550 at which cross-matching of mass spectral components may occur. Eleven groups of components, A1 A11, with average mass spectral match $\geqslant 532$ remained non-identified. An initial manual inspection of the fragmentation pattern and comparison with the commercial NIST98 mass spectra collection allowed no further identification. However, unknown components

A1, A2, A3, A4, A5, and A8 appeared to be similar to quinic acid or caffeoylquinic acids (refer to Section 2.4).

These results strongly support the observation that mass spectral hit lists may reveal previously unknown components from plant tissues which are chemically related to the respective bait substance.

2.3.2. Exemplary search for galactonic acid

In the first application of the MS/RI library caffeoylquinic acid conjugates were chosen because of their highly characteristic fragmentation pattern, which is caused by the presence of two cyclic structural moieties (Fuchs and Spiteller, 1996). In contrast galactonic acid contained only functional groups which are ubiquitous in natural products, for example, carboxyl-groups, primary, and secondary hydroxyl-groups. Furthermore, components exhibiting highly similar mass spectra co-eluted within a small RI window (inset of Fig. 4). Accordingly the range of match factors among the top 400 matching components was only 632 948. Eighteen mass spectral components representing the galatonic acid derivative from plant samples were found and manually confirmed in all but tobacco leaf samples. The average match factor of this group of mass spectral components was 893 ± 79 and $RI = 1991.6 \pm 0.7$ (Table 4). In total, 36 groups of components with group size $\geqslant 3$ were found.

Among the groups of components exhibiting the best mass spectral fit we identified two isomers of galactonic acid, namely the plant products gluconic acid and gulonic acid (Table 4). In addition we found that linear sugar alcohols had high mass spectral match factors = 732 763. Of the reference substances, ribitol, xylitol, mannitol, sorbitol, and galactitol, only ribitol and mannitol were identified within the MS/RI library. Ribitol was an internal standard added to all samples, whereas as mannitol originated from plant samples. Mass spectra of glucose methoxyamine and galactose methoxyamine were also among the high ranking hits and exhibited match factors = 721 746. Methoxyamine moieties are introduced by routine chemical derivatization of carbonyl-groups or cyclic actales and ketales prior to GC/MS profile analysis. Typically a main and a by-product of methoxyamination are found. Finally we identified the plant product threonic acid with match factor $= 754 \pm 25$ within the hit list and detected high similarity to the reference substance glucaric acid, match factor = 743. Fifteen groups of components, B1 B15, remained non-identified. Initial comparison with the commercial NIST98 mass spectra collection allowed no further identification but the components B1, B2, and B3 exhibited match factors $\geqslant 800$ when compared with either gulonic, galactonic, or gluconic acid (see Section 2.4).

Again this case exemplified that unknown components can be identified and that isomers or chemically related compounds can be detected among the top ranking mass spectral components of our MS/RI-

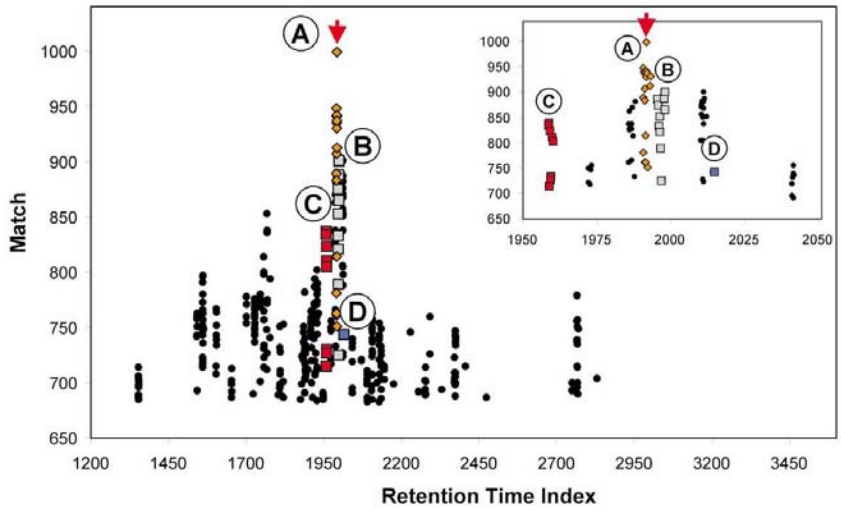


Fig. 4. Chromatographic reconstruction of the mass spectral search list of plant components and reference substances obtained with the bait mass spectrum of persilylated galactonic acid (arrow). The search list contained mass spectra of the siylated derivatives of A (orange diamonds) galactonic acid, B (grey box) gluconic acid, C (red box) gulonic acid, and D (blue box) glucaric acid. Replicate mass spectra from plant samples were confirmed manually. The inset shows RI range 1950—2050.

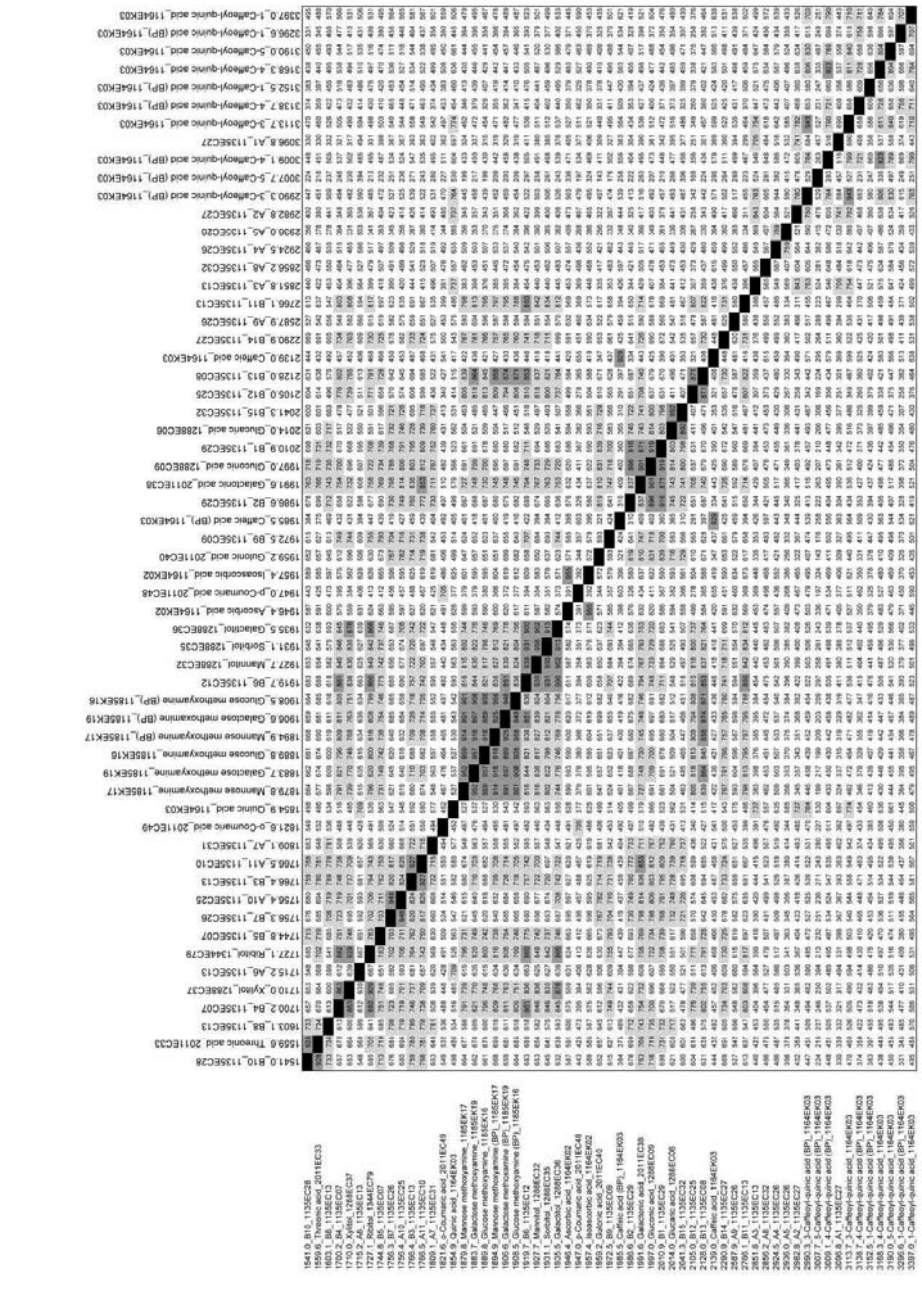


Fig. 5. Complete matrix of pair-wise mass spectral comparisons. Mass spectral components are in chromatographic order. Representative mass spectra of reference substances and unknown components from Tables 3 and 4 (refer to supplemental data 1 and 2) were compared by the NIST98 search algorithm. Match factors describe a range of 1–999, where the maximum value indicates perfect mass spectral identity. Match quality was indicated by grey shades: black, 999 (self matching); dark grey, 850–999; light grey, 700–850.

library. In this example we found in addition to other hexonic acids, $CH_3(HCOH)_4COOH$, compounds like C_5- or C_6-sugar alcohols and methoxyaminated C_6-aldoses, which all contain a straight chain pentitol group, $CH_3(HCOH)_4$-R. Moreover, we identified two matching carboxylic acids with the common structural feature, R-$(HCOH)_3COOH$.

2.4. Classification of unknown mass spectral components

The analysis of the MS/RI library resulted in newly identified substances present within metabolic profiles of plants, however, 11 groups of components with mass spectral similarity to 3-caffeoylquinic acid and 15 groups of components exhibiting similarity to galactonic acid

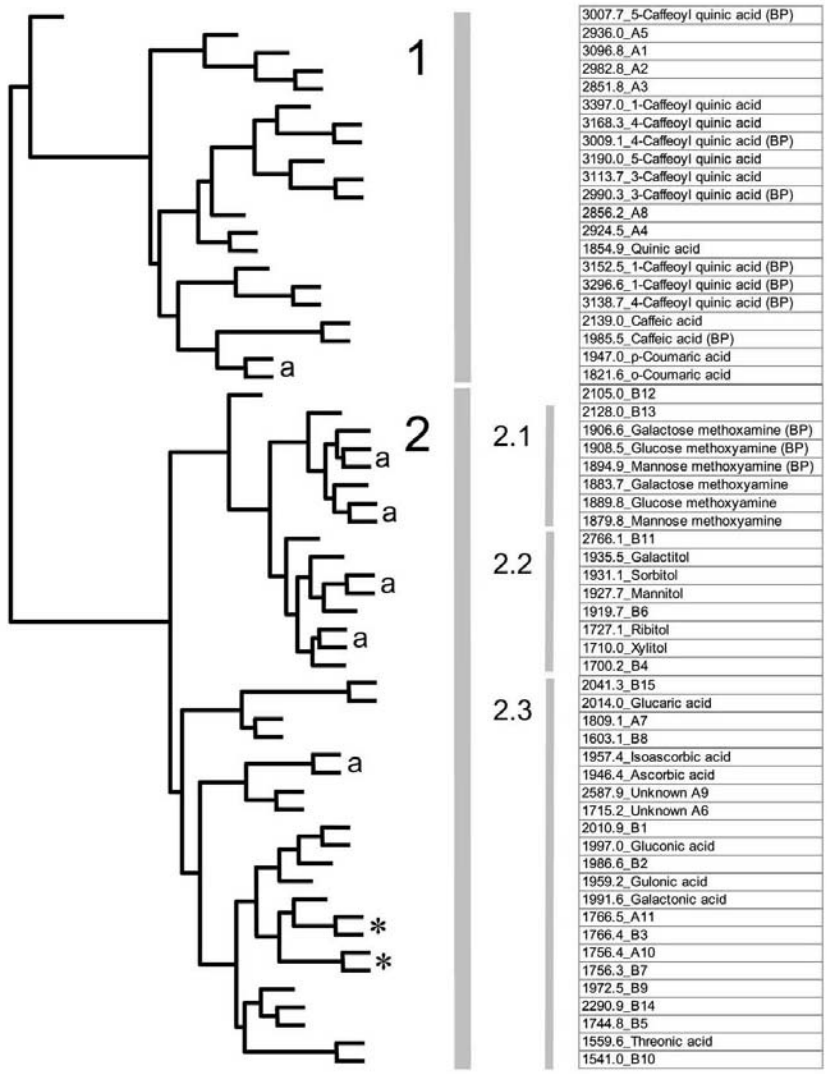


Fig. 6. Hierarchical cluster analysis of mass spectral components based on profiles of mass spectral match factors. The analysis was based on the complete matrix of match factors presented in Fig. 5 (also refer to supplemental data 1). Class 1 comprised cinnamic acid derivatives, quinic acid and respective conjugates. Class 2 was composed of carbohydrates which subgrouped into hexose methoxyamines (2.1), sugar alcohols (2.2), and polyhydroxy-carboxylic acids (2.3). [a] Pairs of conformational isomers. * Redundant mass spectra of identical components.

remained non-identified. We selected representative mass spectra of all groups of unknown components and of available reference substances. The complete set of 59 mass spectra is made available as supplemental file 1 in a data format ready to be imported either into AMDIS software or into the NIST98 mass spectral search program (Appendix). A complete set of 59×59 pair-wise mass spectral comparisons was generated by the NIST98 algorithm (Fig. 5, also see supplemental data 2). Inverse comparisons had identical match factors. The comparison matrix of match factors indicated two main classes of mass spectra with partial chromatographic separation (Fig. 5). A hierarchical cluster analysis based on the comparison matrix of mass spectral match factors generated the tree of nearest neighbours presented in Fig. 6. In contrast to established procedure this classification did not only take best fits into account. We argue that best fits were clearly insufficient for the distinction of groups of redundant components (Tables 3 and 4). In contrast the cluster analysis based on comparison matrices of match factors reflected all available similarities as well as dissimilarities. Among the 20 pairs of nearest neighbours six pairs of conformational isomers and two pairs of identical but unknown components were found, for example, A11/B3, and A10/B7. Identity was confirmed manually by MS and RI comparison of all members of the respective groups from Tables 3 and 4. Furthermore, the resulting tree analysis revealed two main classes of mass spectra (Fig. 6). Class 1 comprised cinnamic acid derivatives, quinic acid and conjugates thereof. Carbohydrates constituted class 2, which was subdivided into hexose methoxyamines, subclass 2.1, sugar alcohols, subclass 2.2, and polyhydroxycarboxylic acids, subclass 2.3, respectively.

Further analysis revealed that the positional isomers of caffeoylquinic acid were aggregated within a single branch of class 1. The 6 non-identified components, A1, A2, A3, A4, A5, and A8, were classified to belong to class 1 and component A4 could be put into close proximity with quinic acid. All other unknown components grouped with class 2 (Fig. 6). In detail, component B6 was suggested to represent a hexitol by proximity to galactitol, sorbitol and mannitol. Component B4 exhibited proximity to pentitols, for example, ribitol and xylitol. Component B15 was closest neighbour of glucaric acid and component B10 matched best with threonic acid. Components B1 and B2 were grouped with the hexonic acids, gluconic and gulonic acid. The component pairs A11/B3 and A10/B7 were also close to hexonic acids as was indicated by occurrence of galactonic acid within the same major branch.

3. Conclusion

Our present work clearly showed the feasibility of non-supervised construction of MS/RI libraries from automatically generated mass spectral components of metabolic profiles. We demonstrated unequivocally that RI was absolutely essential for identification of unknown metabolic components by reference substances and for the grouping of non-identified but redundant mass spectral components. We furthermore demonstrated that selectivity of analysis can be introduced via creation of subsets of mass spectral components which exhibit similarity to single reference mass spectra. Finally we introduced cluster analysis based on matrices of mass spectral match factors and clearly demonstrated applicability of this approach to the classification of unknown mass spectral components. Future efforts will focus on automated error detection of mass spectral deconvolution and the reduction of the initial signal to noise ratio threshold, which currently limits the virtual sensitivity of our MS/RI libraries and thus restricts access to the full sensitivity of GC/EI-TOF-MS (Hirsch et al., 2001; Dalluge et al., 2002). Finally we will attempt a comprehensive classification of full mass spectral libraries comprising non-identified natural products and an extended set of reference substances.

4. Experimental

4.1. Biological materials and sampling

The plant varieties used were potato (*Solanum tuberosum* cv. Désirée), tobacco (*Nicotiana tabaccum* cv. SNN), and *Arabidopsis thaliana* (L.) Heynh., ecotype C24. Potato (*Solanum tuberosum* cv. Désirée) was obtained from Saatzucht Lange AG (Bad Schwartau, Germany). All plants were cultivated on soil in growth chambers with a maximum of 120 μmol photons m^{-2} s^{-1} at leaf surface. Potato and tobacco plants were grown in 3-l pots with a 16 h-light/8 h-dark regime changing from 22 °C during the day to 18 °C at night and with relative humidity preset to constant 70%. *Arabidopsis thaliana* plants were kept in 0.125-l pots with an identical regime of illumination but changing from 60% humidity and 20 °C during the day to 75% humidity and 18 °C at night. Samples of plant organs were harvested during light periods from flowering plants. Root samples were prepared free of soil but not rinsed with water. Plant material was weighed with a precision of ±0.1 mg, immediately frozen in liquid nitrogen, and stored at −70 °C until further analysis. Sampling time before freezing did not exceed 20 s.

4.2. Reference substances and substance identification

L-Threonic acid calcium salt (CAS 70753-61-6) was purchased from Aldrich, Germany; D-(−)-isoascorbic acid (CAS 89-65-6) was from Fluka, Germany;

D-(+)-mannose (CAS 3458-28-4) was ordered from Merck, Germany; L-ascorbic acid (CAS 50-81-7), caffeic acid (CAS 331-39-5), 3-caffeoylquinic acid (CAS 327-97-9), *o*-coumaric acid (CAS 614-60-8), *p*-coumaric acid (CAS 501-98-4), D-(−)-galactonic acid gamma-lactone (CAS 2782-07-2), D-galactonic acid hemicalcium salt (CAS 6622-52-2), D-(+)-galactose (CAS 59-23-4), D-(+)-glucose (CAS 50-99-7), D-glucaric acid monopotassium salt (CAS 576-42-1), D-gluconic acid sodium salt (CAS 527-07-1), D-(+)-gluconic acid delta-lactone (CAS 90-80-2), D-gulonic acid gamma-lactone (CAS 6322-07-2), D-(−)-quinic acid (CAS 77-95-2), ribitol (CAS 488-81-3), and xylitol (CAS 87-99-0), were purchased from Sigma-Aldrich, Germany; D-galactitol (CAS 608-66-2), D-mannitol (CAS 69-65-8), and D-sorbitol (CAS 50-70-4), were from Supelco, Germany.

Solvents and reagents for extraction and derivatization were as follows: methoxyamine hydrochloride was purchased from Sigma-Aldrich, Germany; *N*-methyl-*N*-(trimethylsilyl)-trifluoroacetamide (MSTFA) was from Macherey & Nagel, Germany; pyridine, methanol, and chloroform, all HPLC-grade, were supplied by J.T. Baker (Philipsburg, NJ).

Gulonic acid was not commercially available. This substance was identified as by-product of D-gulonic acid gamma-lactone which was subjected to the methoxyamination and persilylation protocol used for metabolic profiling (see later) and was incubated at room temperature for at least 2 days before GC/MS analysis. In support of this means of identification galactonic acid and gluconic acid were found within profiles generated in parallel from D-(−)-galactonic acid gamma-lactone and D-(+)-gluconic acid delta-lactone, respectively (data not shown).

The isomers of chlorogenic acid (3-caffeoylquinic acid), 4-caffeoylquinic acid, 5-caffeoylquinic acid, and 1-caffeoylquinic acid were synthesized as described earlier (Sefkow, 2001; Sefkow et al., 2001). The final purities of the preparations were at least 90% except for 1-caffeoylquinic acid. By-products occurring within these preparations were used for comparison and classification of components observed within profiles of natural products. Chemical abstracts system (CAS) registry numbers of the reference substances are provided if available.

4.3. GC/EI-TOF-MS metabolite profiling

Extraction, liquid partitioning, concentration to dryness, and methoxyamination of carbonyl-moieties followed by derivatizing acidic protons with *N*-methyl-*N*-(trimethylsilyl)-trifluoroacetamide (MSTFA) prior to GC/EI-TOF-MS analysis was performed essentially as described (Fiehn et al., 2000b). For this study only the polar, methanol/water (1:1; v:v) soluble material of the initial extract was analyzed. Highly lipophilic components were removed by liquid partitioning into chloroform and discarded. The initial protocol was downscaled by a factor of 3.3 in order to accommodate sample loads of 1 60 mg fresh weight. The final reagent volume was 120 µl of pyridine/MSTFA (1:2; v:v). A mixture of the retention time standards, *n*-dodecane (RI 1200), *n*-pentadecane (RI 1500), *n*-nonadecane (RI 1900), *n*-docosane (RI 2200), *n*-octacosane (RI 2800), *n*-dotriacontane (RI 3200), and *n*-hexatriacontane (RI 3600) was included in the final reagent volume. Non-sample control experiments were performed with empty containers used for sampling and storage of plant material. The empty containers were fully processed in parallel with regular samples from extraction to final GC/MS analysis. Non-sample controls were run at the start, initial control, and end, final control, of a series of GC/MS analyses as well as interspersed. Thus initial contamination and GC/MS memory effect were monitored.

A GC 6890 (Agilent Technologies, Palo Alto, CA, USA) was operated under electronic pressure control and equipped with a split/splitless capillary inlet. Injection was 1 µl in the splitless mode with a 2 min pulse at 110 psi and injection temperature set to 230 °C. The capillary column used was a 30 m×0.25 mm inner diameter Rtx-5Sil MS with integrated guard column and a 0.25 µm film (Restek GmbH, Bad Homburg, Germany). Helium was used as carrier gas with constant flow at 1 ml min^{-1}. The temperature program was 2 min at 80 °C followed by a 15 min ramp to 350 °C and final heating for 2 min at 350 °C. The transferline to the mass spectrometer was set to 250 °C.

The time-of-flight (TOF) mass spectrometer was a Pegasus II MS system (Leco, St. Joseph, MI, USA) with an electron impact ionization (EI) source set to 200 °C. Mass spectra were monitored with an acquisition rate of 6 spectra s^{-1} in the mass range $m/z = 70$ 600. Tuning and all other settings of the mass spectrometer were according to manufacturer's recommendations.

4.4. Data processing

Chromatograms were acquired with ChromaTOF™ software (LECO, St. Joseph, MI, USA). Initial processing, namely baseline substraction, smoothing, and export of the processed chromatograms into a *.cdf file interchange format were performed within the ChromaTOF™ software. Retention index calibration and mass spectral deconvolution were performed by AMDIS (Automated Mass Spectral Deconvolution and Identification System, National Institute of Standards and Technology, Gaithersburg, MD, USA) with the settings: adjacent peak subtraction = 2, medium resolution, high sensitivity, and high shape requirements. Those of the resulting components which had at least 0.001% of the total signal of the respective chromatogram were exported into a *.msp library file. The component

names within the *.msp files were edited in order to include sample information which was not automatically exported from AMDIS. The edited component identifier included retention time, retention time index, signal to noise ratio, amount of component, experiment identifier, species, organ, and amount of fresh weight. Finally all edited *.msp files were imported into a combined mass spectral library within the NIST98 mass spectral search program (National Institute of Standards and Technology, Gaithersburg, MD, USA). Mass spectral comparisons were performed in the normal NIST98 identity mode without pre-search or other constraints. Due to the NIST98 software the resulting mass spectral hit lists were restricted to the top 400 best fitting mass spectral components. The hit lists were processed and evaluated by the software package for exploratory data analysis and statistical modelling, S-Plus 2000 standard edition release 3 (Insightful, Berlin, Germany).

Data processing with the aim to classify unknown mass spectral components was performed as follows. A mass spectral library of 59 representative identified and unknown components was compiled from GC/EI-TOF-MS profiles of reference substances and plant matrices. Each of the single components was compared separately with the complete library. The 59 resulting mass spectral hit lists were combined into a single 59×59 comparison matrix (Fig. 5) using the match factor as sole measure of mass spectral similarity. In addition to the mass spectral match factor the NIST98 software reported a reverse match factor and a probability measure of mass spectral identity. We compared the respective values of initial and inverse mass spectral comparisons of all pairs of mass spectra. Only the match factor exhibited identical values for all bi-directional, pair-wise comparisons. The reverse match factor and probability measure were disregarded for further analysis. The hierarchical cluster analysis of the comparison matrix was agglomerative using the euclidian dissimilarity measure and average linkage.

Acknowledgements

The authors thank Dr. Alisdair Fernie and Professor Lothar Willmitzer, Max-Planck-Institute of Molecular Plant Physiology, Golm, Germany, for valuable discussions, editorial suggestions and support of our work.

Appendix. Supplementary data

Supplementary file 1

Data sheet constituting Fig. 5 and used as basis for the hierarchical cluster analysis of mass spectra which is presented in Fig. 6.

The data set describes the complete matrix of pair-wise mass spectral comparisons generated by the NIST98 search algorithm. Mass spectral components are in chromatographic order on both axes. Match factors describe a range of 1 999, where the maximum value indicates perfect mass spectral identity. Match quality was indicated by colour: black, 999 (self matching); orange, 850 999; yellow, 700 - 850. Representative mass spectra of reference substances and unknown components from Tables 3 and 4 may be downloaded from supplemental file 2.

Supplementary file 2

The datafile, MS_RI.msp[1] contains representative GC/EI-TOF-MS mass spectra of all reference substances and unknown components mentioned within Tables 3 and 4.

The spectrum name was designed to allow sorting according to retention time index, e.g. 1766.4_B3_1135EC13 or 3113.7_3-Caffeoylquinic acid_1164EK03. The name of the mass spectrum contains three types of information separated by (_). The first position denotes retention time index, the second position indicates name of reference substance[2] or unknown component (refer to Tables 3 and 4), third position encodes the experiment identifier.[3]

References

Baumert, A., Mock, H.-P., Schmidt, J., Herbers, K., Sonnewald, U., Strack, D., 2001. Patterns of phenylpropanoids in non-inoculated and potato virus Y-inoculated leaves of transgenic tobacco plants expressing yeast-derived invertase. Phytochemistry 56, 535–541.

Boros, L.G., Cascante, M., Lee, W.N.P., 2002. Metabolic profiling of cell growth and death in cancer: applications in drug discovery. Drug Discovery Today 7, 364–372.

Dalluge, J., Vreuls, R.J.J., Beens, J., Brinkman, U.A.T., 2002. Opimization and characterization of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometric detection (GC×GC-TOF-MS). Journal of Separation Science 25, 201–214.

Duez, P., Kumps, A., Mardens, Y., 1996. GC-MS profiling of urinary organic acids evaluated as a quantitative method. Clinical Chemistry 42, 1609–1615.

Dyson, N., 1999. Peak distortion, data sampling errors and the integrator in the measurement of very narrow chromatographic peaks. Journal of Chromatography A 842, 321–340.

[1] The file format *.msp can be imported into NIST98 mass spectral comparison software (to be downloaded from http://chemdata.nist.gov/mass-spc/Srch_v1.7/index.html or AMDIS software (to be downloaded from http://chemdata.nist.gov/mass-spc/amdis/).

[2] By-products observed in preparations of reference substances were marked (BP).

[3] Experiments representing metabolic profiles of plant samples were listed in Table 1. Experiments with reference substances were indicated within the mass spectral name but were not further mentioned in this work.

Fan, T.W.-M., Lane, A.N., Shenker, M., Bartley, J.P., Crowley, D., Higashi, R.M., 2001. Comprehensive chemical profiling of gramineous plant root exudates using high-resolution NMR and MS. Phytochemistry 57, 209–221.

Fiehn, O., 2002. Metabolomics—the link between genotypes and phenotypes. Plant Molecular Biology 48, 155–171.

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R.N., Willmitzer, L., 2000a. Metabolite profiling for plant functional genomics. Nature Biotechnology 18, 1157–1161.

Fiehn, O., Kopka, J., Trethewey, R.N., Willmitzer, L., 2000b. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. Analytical Chemistry 72, 3573–3580.

Frenzel, T., Miller, A., Engel, K.H., 2002. Metabolite profiling—a fractionation method for analysis of major and minor compounds in rice grains. Cereal Chemistry 79, 215–221.

Fuchs, C., Spiteller, G., 1996. Rapid and easy identification of isomers of coumaroyl- and caffeoyl-D-quinic acid by gas chromatography mass spectrometry. Journal of Mass Spectrometry 31, 602–608.

Griffin, J.L., Williams, H.J., Sang, E., Clarke, K., Rae, C., Nicholson, J.K., 2001. Metabolic profiling of genetic disorders: a multitissue H-1 nuclear magnetic resonance spectroscopic and pattern recognition study into dystrophic disorders. Analytical Biochemistry 293, 16–21.

Griffiths, D.W., Bain, H., 1997. Photo-induced changes in the concentrations of individual chlorogenic acid isomers in potato (*Solanum tuberosum*) tubers and their complexation with ferric ions. Potato Research 40, 307–315.

Hirsch, R., Ternes, T.A., Bobeldijk, I., Weck, R.A., 2001. Determination of environmentally relevant compounds using fast GC/TOF-MS. Chimia 55, 19–22.

Huhman, D.V., Sumner, L.W., 2002. Metabolic profiling of saponins in *Medicago sativa* and *Medicago trunculata* using HPLC coupled to an electrospray ion-trap mass spectrometer. Phytochemistry 59, 347–360.

Koeppe, D.E., Rohrbaugh, L.M., Wender, S.H., 1969. The effect of varying u.v. intensities on the concentration of scopolin and caffeoylquinic acids in tobacco and sunflower. Phytochemistry 8, 889–896.

Kuhara, T., 2001. Diagnosis of inborn errors of metabolism using filter paper urine, urease treatment, isotope dilution and gas chromatography–mass spectrometry. Journal of Chromatography B 758, 3–25.

Matsumoto, I., Kuhara, T., 1996. A new chemical diagnostic method for inborn errors of metabolism by mass spectrometry—rapid, practical, and simultaneous urinary metabolites analysis. Mass Spectrometry Reviews 15, 43–57.

Percival, G.C., Baird, L., 2000. Influence of storage upon light-induced chlorogenic acid accumulation in potato tubers (*Solanum tuberosum* L.). Journal of Agricultural and Food Chemistry 48, 2476–2482.

Sefkow, M., 2001. First efficient synthesis of chlorogenic acid. European Journal of Organic Chemistry 6, 1137–1141.

Sefkow, M., Kelling, A., Schilde, U., 2001. First efficient syntheses of 1-, 4-, and 5-caffeoylquinic acid. European Journal of Organic Chemistry 14, 2735–2742.

Roessner, U., Wagner, C., Kopka, J., Trethewey, R.N., Willmitzer, L., 2000. Simultaneous analysis of metabolites in potato tuber by gas chromatography–mass spectrometry. Plant Journal 23, 131–142.

Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L., Fernie, A.R., 2001a. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. Plant Cell 13, 11–29.

Roessner, U., Willmitzer, L., Fernie, A.R., 2001b. High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. Plant Physiology 127, 749–764.

van Deursen, M.M., Beens, J., Janssen, H.-G., Leclercq, P.A., Cramers, C.A., 2000. Evaluation of time-of-flight mass spectrometric detection for fast gas chromatography. Journal of Chromatography A 878, 205–213.

Veriotti, T., Sacks, R., 2000. High speed GC/MS of gasoline-range hydrocarbon compounds using a pressure-tuneable column ensemble and time-of-flight detection. Analytical Chemistry 72, 3063–3069.

Veriotti, T., Sacks, R., 2001. High-speed GC and GC/ time-of-flight MS of lemon and lime oil samples. Analytical Chemistry 73, 4395–4402.

BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 16 2004, pages 2841–2844
doi:10.1093/bioinformatics/bth278

PaVESy: Pathway Visualization and Editing System

Alexander Lüdemann, Daniel Weicht, Joachim Selbig and Joachim Kopka*

Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm, Germany

Received on January 23, 2004; revised on March 23, 2004; accepted on April 7, 2004
Advance Access publication April 22, 2004

ABSTRACT

Summary: A data managing system for editing and visualization of biological pathways is presented. The main component of PaVESy (Pathway Visualization and Editing System) is a relational SQL database system. The database design allows storage of biological objects, such as metabolites, proteins, genes and respective relations, which are required to assemble metabolic and regulatory biological interactions. The database model accommodates highly flexible annotation of biological objects by user-defined attributes. In addition, specific roles of objects are derived from these attributes in the context of user-defined interactions, e.g. in the course of pathway generation or during editing of the database content. Furthermore, the user may organize and arrange the database content within a folder structure and is free to group and annotate database objects of interest within customizable subsets. Thus, we allow an individualized view on the database content and facilitate user customization. A JAVA-based class library was developed, which serves as the database programming interface to PaVESy. This API provides classes, which implement the concepts of object persistence in SQL databases, such as entries, interactions, annotations, folders and subsets. We created editing and visualization tools for navigation in and visualization of the database content. User approved pathway assemblies are stored and may be retrieved for continued modification, annotation and export. Data export is interfaced with a range of network visualization programs, such as Pajek or other software allowing import of SBML or GML data format.

Availability: http://pavesy.mpimp-golm.mpg.de/

Contact: luedemann@mpimp-golm.mpg.de

INTRODUCTION

Multi-parallel technologies, which allow joined analysis of the transcriptome, the proteome and especially the metabolome of organisms (Fiehn *et al.*, 2000) today allow the investigation of complex biological processes at full systems level. The tools of modern functional genomics provide means to systematically modify gene expression of each gene present within an organism and to monitor changes in gene expression, enzyme activity and metabolite levels. The major task of the time to come will be the discovery of novel functional interactions within and in between the metabolite, protein and mRNA complement of biological systems (Urbanczyk-Wochniak *et al.*, 2003). Novel findings need to be mapped onto metabolic pathways and compared with the long-standing knowledge of biochemical reactions and regulatory interactions. The ultimate goal will be to fully understand the function of each gene, encoded protein and effected metabolite.

Currently, public knowledge about biochemical reactions, pathways and underlying genes is deposited within a range of in part complementary, in part competing databases. Publicly accessible databases, e.g. the GenBank (http://www.ncbi.nlm.nih.gov/entrez), the KEGG system (Goto *et al.*, 2002), the BRENDA enzyme database (Schomburg *et al.*, 2002), the EcoCyc/MetaCyc system (Karp *et al.*, 2002), WIT (Overbeek *et al.*, 2000), the BIND (Bader *et al.*, 2000), GEO (Edgar *et al.*, 2002) and ArrayExpress (Brazma *et al.*, 2003) database or the MPW project (Selkov *et al.*, 1998) harbour data, which cannot be edited, updated or manipulated according to the needs of expert scientists. In contrast to these static databases, which provide predefined reactions and fixed metabolic maps, more recent developments already allow dynamic retrieval and visualization of metabolic pathways, e.g. the commercial PathBlazer tool (http://informaxinc.com/content.cfm?pageid=17) or public domain systems (Krishnamurthy *et al.*, 2003; Pan *et al.*, 2003; Trost *et al.*, 2003), such as Cytoscape (Shannon *et al.*, 2003), Osprey (Breitkreutz *et al.*, 2002) or TopNet (Yu *et al.*, 2004). However, user annotation is in general restricted to predefined sets of attributes for each biological object and to a fixed number of possible interactions. In addition, calculations aimed at analysis and comparison of network topology and metabolic connectivity still require downloading of available metabolic data sets from public sources into a customized database.

*To whom correspondence should be addressed.

We developed PaVESy (Pathway Visualization and Editing System) to integrate and compare numeric multi-parallel genomic, proteomic and metabolomic data with respect to publicly available pathway information and expert knowledge, which can be embedded by customized definitions of objects, attributes and interactions. The design was made for the purpose of data mining in the fields of metabolic profiling, metabolomics and plant functional genomics, but system architecture remained open for other applications.

SYSTEM OVERVIEW

PaVESy is a three-layered tool comprising a relational database, a data object layer and a graphical user interface, which allow editing, storage and retrieval of biological objects and interactions, assembly of metabolic pathways and export to visualization tools for pathway reconstruction. Data object layer and graphical user interface are implemented in a JAVA class package, which guarantees object persistence at SQL level, comprises a transaction concept and among others provides design patterns of attribute holders, biological objects and metabolic or regulatory interactions (see below).

The PaVESy database model

The PaVESy database comprises cellular components, such as metabolites, proteins/enzymes and genes, which are required to reconstruct metabolic and regulatory networks (Ravasz *et al.*, 2002). Cellular components form modules by multiple types of interactions. In the following, first level interactions represent direct relations of two or more biological components, e.g. gene-encodes-protein, protein-binds-gene (transcription factors), protein-binds-ligand, protein-binds-protein, metabolic effector interactions or biochemical reactions. In contrast, second level interactions describe composite interactions, such as metabolic or regulatory pathways assembled from a selection of first level interactions, e.g. biochemical reactions or regulatory interactions.

The PaVESy design accommodates the properties of biological objects and interactions within five database tables, namely the ENTRY, ENTITY, ARC, ATTRIBUTE and ATTRIBUTE_DEF tables. The design of ENTRY, ENTITY and ARC tables is similar to the PathBlazer tool (InforMax Inc., Oxford, UK). In short, the ENTRY table harbours unique database accessions of all database objects, biological components, interactions, pathways and container objects, such as subsets and folders. The ENTITY table records copies of objects from the ENTRY table with membership to an interaction and role in the context of this interaction. The ARC table records the connectivity of biological objects and directionality of interactions for network reconstruction. Each record of the ENTRY, ENTITY and ARC tables can be specifically attributed by records of the ATTRIBUTE table. The meta-information of the attributes, e.g. definition of type, property, number and occurrence, are defined by entries in the ATTRIBUTE_DEF table. The annotation of database objects can thus be easily queried and dynamically customized without rebuilding the database structure. The database is non-normalized. Instead the structure of database tables was optimized to avoid table joining for enhanced search and pathway reconstruction algorithms.

The PaVESy explorer

The PaVESy explorer window (Fig. 1A) allows intuitive navigation. The panel to the left displays database objects in a tree-like folder structure, which also allows to organize and assemble subsets of customized metabolic and regulatory pathways. The panel to the right shows object attributes and allows expert annotation and editing. Assembly of information from the database into customized subsets is aided by search windows, such as the compound search (Fig. 1D). Search protocols implement SQL-based options for wild cards. This feature is highly required in view of the presence of multiple synonyms for biological objects.

Pathway assembly

Pathways can be assembled either from sets of pre-selected reactions or by automated retrieval from the database. A pathway generation window serves these purposes (Fig. 1C). Assembly without pre-selection of reactions requires the definition of a start and an end component, as well as the definition of maximum step length and generality. A combination of depth-first and breadth-first graph search algorithm was used (Krishnamurthy *et al.*, 2003). Sets of components, such as water, ATP, ADP, P_i or other ubiquitous cofactors, which are ignored during automated pathway assembly can be defined. Visualization of pathways comprising components with high network connectivity is highly facilitated by optional use of subsets of 'non-pooled' items. These components will not be pooled into unique vertices. An example of pathway assembly using citric acid and isocitric acid as anchor components and defining aconitate hydratase to be non-pooled is shown in Figure 1B. Runtime for retrieving this examplary pathway from a download of the KEGG data set was <0.5 s. The results of the pathway generation can be stored for further use or exported into different file formats. Currently supported formats are the Pajek net format (Batagelj and Mrvar, 2004), SBML (Hucka *et al.*, 2003, http://www.sbml.org) and GML (http://infosun.fmi.uni-passau.de/Graphlet/GML/). Pajek software is publicly available and provides tools and layout algorithms for analyzing large networks. SBML and GML formats can be uploaded into public domain Cytoscape software.

The PaVESy program may be downloaded from http://pavesy.mpimp-golm.mpg.de/ and runs under MS-Windows, MAC-OS or Linux platforms. The Java runtime environment (JRE) Version 1.4 or higher and a SQL database, such as MS-Access or SAP-DB (http://www.sapdb.org/) must be installed.

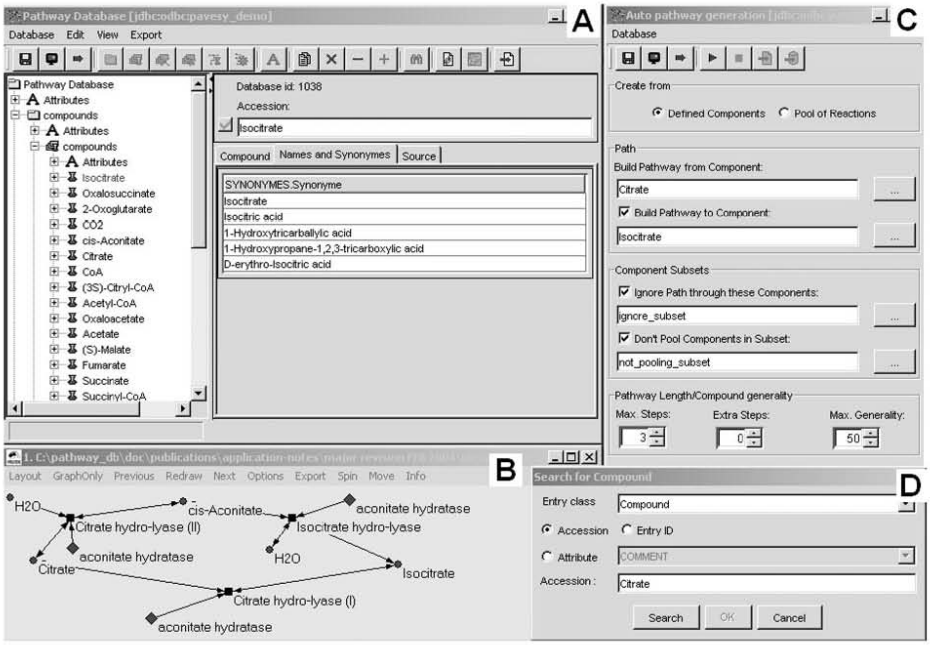


Fig. 1. Example of the PaVESy explorer window and visualization of pathways by Pajek software. The panel to the left displays a tree-like folder structure of the database content (**A**) and the panel to the right shows the object attributes and allows editing. The Pajek visualization is displayed in (**B**) and the pathway generation window (**C**) displaying an exemplary search for reactions interconnecting citric acid and isocitric acid. (**D**) Shows the database search window.

FUTURE WORK

For the future, it is intended to integrate an import and export interface for up- and downloading data. The structure of the data should correspond to SBML Level 2, so that it is possible to use different tools that support this SBML definition. Another important point is the integration of different database systems. The user should be able to choose the system he/she prefers to work with just by choosing the database driver. To complete PaVESy, an own pathway visualization application should be integrated.

ACKNOWLEDGEMENTS

This work was supported by the Max Planck society. The authors thank Prof. L. Willmitzer and Prof. M. Stitt for encouragement, continued support and discussion of this work. We are grateful to Prof. D. Schomburg for allowing access to the BRENDA collection (http://www.brenda.uni-koeln.de) and acknowledge the Kyoto Encyclopedia of Genes and Genomes (http://www.genome.ad.jp/kegg) for providing public access to metabolic data and reactions.

REFERENCES

Bader,G.D., Donaldson,I., Wolting,C., Ouellette,B.F.F., Pawson,T. and Hogue,C.W.V. (2001) BIND—the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **29**, 242–245.

Batagelj,V. and Mrvar,A. (2004) Graph drawing software: Pajek—analysis and visualization of large networks. In Junger,M. and Mutzel,P. (eds), *Mathematics and Visualization*. Springer, Berlin, pp. 77–103.

Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P., Lara,G.G. *et al.* (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **32**, 68–71.

Breitkreutz,B.-J., Stark,C. and Tyers,M. (2002) Osprey: a network visualization system. *Genome Biol.*, **3**, PREPRINT0012.

Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Fiehn,O., Kopka,J., Dormann,P., Altmann,T., Trethewey,R.N. and Willmitzer,L. (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.*, **18**, 1157–1161.

Goto,S., Okuno,Y., Hattori,M., Nishioka,T. and Kanehisa,M. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.*, **30**, 402–404.

Hucka,M., Finney,A., Sauro,H.M., Bolouri,H., Doyle,J.C., Kitano,H., Arkin,A.P., Bornstein,B.J., Bray,D., Cornish-Bowden,A. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.

Karp,P.D., Riley,M., Paley,S.M. and Pellegrini-Toole,A. (2002) The MetaCyc database. *Nucleic Acids Res.*, **30**, 59–61.

Krishnamurthy,L., Nadeau,J., Ozsoyoglu,G., Ozsoyoglu,M., Schaeffer,G., Tasan,M. and Xu,W. (2003) Pathways database system: an integrated system for biological pathways. *Bioinformatics*, **19**, 930–937.

Overbeek,R., Larsen,N., Pusch,G.D., D'Souza,M., Selkov,E.,Jr, Kyrpides,N., Fonstein,M., Maltsev,N. and Selkov,E. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.*, **28**, 123–125.

Pan,D., Sun,N., Cheung,K.-H., Guan,Z., Ma,L., Holford,M., Deng,X.W. and Zhao,H. (2003) PathMAPA: a tool for displaying gene expression and performing statistical tests on metabolic pathways at multiple levels for Arabidopsis. *BMC Bioinformatics*, **4**, 56.

Ravasz,E., Somera,A.L., Mongru,D.A., Oltvai,Z.N. and Barabási,A.-L. (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Schomburg,I., Chang,A. and Schomburg,D. (2002) BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.*, **30**, 47–49.

Selkov,E., Grechkin,Y., Mikhailova,N. and Selkov,E. (1998) MPW: the metabolic pathways database. *Nucleic Acids Res.*, **26**, 43–45.

Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Trost,E., Hackl,H., Maurer,M. and Trajanoski,Z. (2003) Java editor for biological pathways. *Bioinformatics*, **19**, 786–787.

Urbanczyk-Wochniak,E., Luedemann,A., Kopka,J., Selbig,J., Roessner-Tunali,U., Willmitzer,L. and Fernie,A.R. (2003) Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep.*, **4**, 989–993.

Yu,H., Zhu,X., Greenbaum,J.K. and Gerstein,M. (2004) TopNet: a tool for comparing biological subnetworks, correlating protein properties with topological statistics. *Nucleic Acids Res.*, **32**, 328–337.

FEBS Letters 579 (2005) 1332 1337 FEBS 29286

Hypothesis

GC–MS libraries for the rapid identification of metabolites in complex biological samples

Nicolas Schauer[a], Dirk Steinhauser[a], Sergej Strelkov[b], Dietmar Schomburg[b], Gordon Allison[c], Thomas Moritz[d], Krister Lundgren[d], Ute Roessner-Tunali[e], Megan G. Forbes[e], Lothar Willmitzer[a], Alisdair R. Fernie[a], Joachim Kopka[a,*]

[a] *Max-Planck Institute of Plant Molecular Physiology, Am Muehlenberg 1, D-14476 Golm, Germany*
[b] *University of Köln, CUBIC Institute of Biochemistry, Zuelpicher Street 47, D-50674 Köln, Germany*
[c] *Institute of Grassland and Enviromental Research, SY23 3EB, Aberystwyth, Wales, UK*
[d] *Umea Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, SE-901 83 Umea, Sweden*
[e] *Australian Centre for Plant Functional Genomics, School of Botany, University of Melbourne, Vic. 3010, Australia*

Received 17 December 2004; revised 12 January 2005; accepted 19 January 2005

Available online 28 January 2005

Edited by Lukas Huber

Abstract Gas chromatography–mass spectrometry based metabolite profiling of biological samples is rapidly becoming one of the cornerstones of functional genomics and systems biology. Thus, the technology needs to be available to many laboratories and open exchange of information is required such as those achieved for transcript and protein data. The key-step in metabolite profiling is the unambiguous identification of metabolites in highly complex metabolite preparations with composite structure. Collections of mass spectra, which comprise frequently observed identified and non-identified metabolites, represent the most effective means to pool the identification efforts currently performed in many laboratories around the world. Here, we describe a platform for mass spectral and retention time index libraries that will enable this process (MSRI; www.csbdb.mpimp-golm.mpg.de/gmd.html). This resource should ameliorate many of the problems that each laboratory will face both for the initial establishment of metabolome analysis and for its maintenance at a constant sample throughput.

Keywords: GC–MS; Gas chromatography–mass spectrometry; Mass spectral library; Metabolite profiling; Metabolomics; Retention time index

1. Introduction

In the last decade, the maturity of genomic technologies generated a vast amount of sequence data and thus allowed full insight into the finite number of genes which constitute organisms. As a consequence, biological science went through a paradigm-change and today focuses on unravelling gene function and regulation. With these tasks at hand, comprehensive technologies have been developed which aim at comprehensive and non-biased monitoring of gene expression, and coinciding effects on protein composition and changes in metabolism. Consequently, new fields emerged in biological science which we today call transcriptomics, proteomics and metabolomics. With increasing amount and diversity of "-omics" data, the need for standardization by the research community arises and availability of tools for a user-friendly, open access to the flood of information has become essential.

One of the first "-omics" databases, BRENDA, was developed in 1987 [1]. BRENDA is a powerful database of enzyme and metabolic information initially published as a series of books, now adapted to a relational database and accessible through the worldwide-web. BRENDA hosts about 83 000 different enzymes from 9800 different organisms and describes enzyme function, taxonomy, sequences and enzyme ligands. The Munich Information Centre for Protein Sequences (MIPS) provides databases related to protein sequences based on whole genome analysis and annotation. For example, MIPS hosts databases of *Saccharomyces cerevisiae* and *Neurospora crassa* which comprise maps of protein–protein interactions, protein localization, and information on transcription factors, cDNA libraries and gene homology [2]. Transcript profiling rapidly evolved into a worldwide accepted and generally applied laboratory tool. Subsequently, databases were designed and established, which efficiently deal with transcriptome data. The Stanford microarray database (SMD), which hosts data of over 3500 DNA-microarrays of 12 distinct organisms, including bacteria, plants and animals, was the first implementation to fulfil this aim [3].

With the full availability of the human genome sequence the need and opportunity to understand the structure and function of all proteins, beyond those with enzymatic properties, was met with respective initiatives. For example HPI, the human protein initiative, focuses on the annotation of both the human genome and proteome. As proteins are generally regarded to determine cellular function, the full exploration of the prote-

*Corresponding author.
E-mail address: kopka@mpimp-golm.mpg.de (J. Kopka).

Abbreviations: GC, gas chromatography; MS, mass spectrometry; MST, mass spectral metabolite tag; RI, retention time index; TOF, time of flight

0014-5793/$30.00
doi:10.1016/j.febslet.2005.01.029

ome will be crucial. The goal of HPI is to deliver this information in high quality to facilitate further investigations of the genomic and proteomic data [4].

Presently, a wealth of databases houses information gathered at the genomic, transcriptomic, proteomic and metabolomic levels of live, e.g. [1,5]. However, there is a significant lack of a metabolome database, capable of storing the flood of data arising from analysis of biological samples using established gas chromatography–mass spectrometry (GC–MS) techniques for metabolome [6–9] and fluxome analysis [10–12]. Most promisingly, first efforts have already been made by the plant metabolomics community to agree on conventions for data formats and the description of metabolomics experiments [13,14].

2. GC–MS based metabolome analysis: Application and key challenge

GC–MS based metabolome analysis has profound applications in discovering the mode of action of drugs or herbicides and helps unravel the effect of altered gene expression on metabolism and organism performance in biotechnological applications. The prerequisite and thus key challenge of metabolite profiling is the rapid, reliable and unambiguous identification of hundreds of metabolites in highly complex preparations, such as blood plasma, intracellular microbial extracts, or complex plant and animal samples. Identification is routinely performed by time-consuming standard addition experiments using commercially available or purified metabolite preparations. Thus, a strong need for a publicly accessible database exists, harbouring the evidence and underlying metabolite identification in complex GC–MS profiles from diverse biological sources. In addition, the non-supervised collection of as yet unidentified mass spectra of metabolites, "so-called" mass spectral metabolite tags (MSTs), will most likely be highly effective for future identification efforts and discovery of novel metabolic markers. In this report, we present a platform of mass spectral and retention time index (MSRI) libraries, generated using identical types of capillary GC columns, however, utilizing two independent GC–MS detection technologies, namely quadrupole (QUAD) GC–MS [6,7,9,15] and GC-TOF (time of flight)-MS [8,16]. In the following study, we will present three test cases which illustrate the general applicability of this library for the key processes of GC–MS based metabolite profiling, (i) identification or preliminary classification of all MST components, which are present in any given biological sample, (ii) query for those biological samples that contain a certain metabolite, (iii) matching of metabolite identifications made on different GC–MS systems and by different laboratories.

3. Mass spectral and retention time index libraries for GC–MS

We propose public exchange and open access of mass spectral identifications from GC–MS metabolite profiles, for example, through a web-based platform of MSRI libraries (www.csbdb.mpimp-golm.mpg.de/gmd.html [17]). In addition, we provide downloadable files, which can be imported into the currently leading and widely accepted NIST02 mass spectral search program or AMDIS, the automated mass spectral deconvolution and identification system (National Institute of Standards and Technology, Gaithersburg, MD, USA) [18,19]. Both software systems are publicly available from www.chemdata.nist.gov/mass-spc/amdis/ and www.chemdata.nist.gov/mass-spc/Srch_v1.7/index.html. Our libraries are classified according to technology and degree of manual mass spectral identification that was required for the library construction. After import into NIST02, the current libraries may be fused into one or customized subsets generated. Q_MSRI and T_MSRI libraries contain MSTs, which were either generated on three identically configured quadrupole (Q_MSRI) GC–MS systems or on a single time of flight (T_MSRI) system. All systems were run with identical settings except for the temperature program and scanning rate (refer to the MSRI Library: Methods on the web). Mass spectral libraries, which exclusively comprise manually evaluated, identified or classified MSTs, are assigned to ID-libraries, indicative of supervised identifications. Libraries which were generated exclusively by automated deconvolution were assigned NS indicative of the non-supervised mode of construction. The NS-libraries may contain deconvolution errors, such as multiple mass spectra for single components, accidental deconvolutions, due to random fluctuations of background noise, or partial and mixed, in other words, chimeric mass spectra of metabolic components. In addition, detailed information on processed biological samples, source of pure reference compounds, respective collaborators and previous citations is provided. For those queries on the current mass spectral

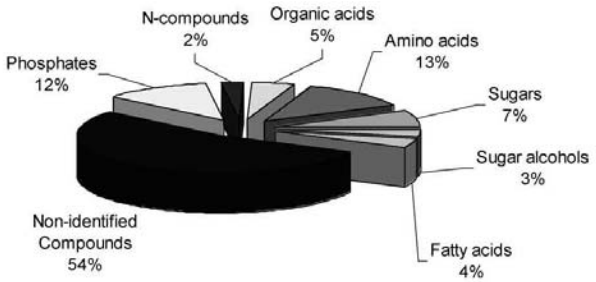


Fig. 1. The distribution of major compound classes, which were identified in GC-TOF-MS profiles in cellular extracts from *Corynebacterium glutamicum*. Note that more than 50% of the metabolites that are covered by GC MS are currently non-identified MSTs.

Table 1
Metabolites found in preparations of blood plasma samples from domestic sheep

Class	Metabolite
Amino acids[a]	2-Aminobutyric acid
	4-Hydroxyproline
	Alanine
	β-Alanine
	Glycine
	Alanine
	Arginine
	Asparagine
	Cysteine
	Glutamic acid
	Glutamine
	Homoserine
	Isoleucine
	Leucine
	Lysine
	Methionine
	Phenylalanine
	Proline
	Serine
	Threonine
	Tryptophan
	Tyrosine
	Valine
	N-Acetylglycine
	Ornithine
	Pyroglutamic acid
	S-Methyl-cysteine
Organic acids	2-Ketoglutaric acid
	4-Hydroxybenzoic acid
	Benzoic acid
	Citric acid
	Erythronic acid
	Fumaric acid
	Gluconic acid
	Glucuronic acid
	Glutaric acid
	Glyceric acid
	Gulonic acid
	Isocitric acid
	Itaconic acid
	Malic acid
	Threonic acid
	trans-Sinapinic acid
Lipids	9,12-(*cis*,*cis*)-Octadecadienoic acid
	Hexadecanoic acid
	Octadecanoic acid
	α-Tocopherol
	β-Sitosterol
	Campesterol
	Cholesterol
Phosphates	Adenosine-5-monophosphate
	Glycerol-3-phosphate
	myo-Inositol-phosphate
	Phosphoric acid
Sugars[a]	Arabinose
	Fructose
	Glucose
	Raffinose
	Ribose
N-compounds	Allantoin
	Hypoxanthine
	Inosine
	Thymine
Alcohols	Erythritol
	Glycerol
	myo-Inositol
	Sorbitol
	Threitol
	Xylitol
	Kaempferol

[a]With rare exceptions DL stereoisomers are not separated on the current choice of GC capillary column.

collection, which cannot be performed within NIST02, we offer a tab delimited compilation of the manually evaluated mass spectra (refer to the MSRI library: descriptions on the web) and access through web query forms. Currently, we support queries within ID-libraries, such as compound search, mass spectral search using names or mass spectra and customized library generation for subsets of mass spectra.

We previously demonstrated that both mass spectrum and retention time index are required for unequivocal metabolite identification in GC–MS profiles [16]. This feature was not available in commercial mass spectral comparison software. Therefore, the central feature of our web search forms is optional restriction of searches to RI windows and sorting of hit lists according to RI deviation and mass spectral similarity. For a shortlist of the currently implemented matching tools and queries please refer to [17].

The present version of the Q_MSRI_ID library contains 1166 identified or annotated MSTs, which represent 574 non-redundant compounds. Of these compounds 306 are unambiguously identified, while the residual MSTs are annotated with the best mass spectral match from a commercially available mass spectral collection (National Institute of Standards and Technology, Gaithersburg, MD, USA). The T_MSRI_ID collection has a similar size, namely 855 MSTs with 229 identifications within the set of 632 non-redundant components. The non-supervised collections comprise close to 30000 MSTs from a range of plant organs, root, leaf, tuber, stolon, flower, and fruits in different developmental stages, and suitable non-samples controls. Plant species covered are model plants, crops and related wild species, such as *Lotus japonicus*, *Arabidopsis thaliana*, *Solanum tuberosum*, *Nicotiana tabacum*, *Solanum lycopersicum*, *Solanum pennellii*, *Solanum parviflorum*, *Solanum pimpinellifolium*, *Solanum habrochaites*, *Solanum neorickii*.

4. Test cases

4.1. Test case 1: Analysis of sample composition

Non-supervised MSRI data allow screening for differences in various samples. A given biological sample can be compared to the non-supervised library. All MSTs, which match in their mass spectrum and RI, within certain thresholds, such as mass spectral match >650 and RI deviation <3.0, will be presented as possible hits, thus allowing the evaluation of whole biological samples for differences in composition with respect to mass spectral datasets from the established MSRI library. In the following, we applied the above thresholds for automated identification but still performed additional manual verification on each of the best hits.

A supervised database is a valuable tool to identify compounds with known RI and mass spectra in specific biological samples. A typical example of the metabolite composition from polar bacterial extracts demonstrates the scope of GC–MS based metabolite profiling (Fig. 1). To further illustrate the power of this tool, we have chosen the plant specific flavonol (kaempferol), a phytosterol (β-sitosterol) and vitamin E (α-tocopherol). Taking into account that sheep are herbivores, we expected to find β-sitosterol and kaempferol also in sheep plasma samples. To test our hypothesis, we have performed a MSRI library searching for these compounds in sheep blood plasma samples, resulting in mass spectral hits for β-sitosterol and kaempferol in the plasma composition (Table 1).

It is also known that tocopherol and its derivatives play an important role in the human diet and thus are important targets for novel nutrigenomics approaches [20,21]. Tocopherol is additionally widely hypothesized to be helpful in preventing diseases associated with oxidative stress. Therefore, the question arose whether this substance can be easily identified in mammal tissues? Here, we demonstrate the power of a MSRI library to search for α-tocopherol in different samples from animal, as well as plant tissues. Considering the importance of this compound for mammals, we expected to find α-tocopherol in animal samples and querying the library indeed resulted in the identification of α-tocopherol in blood plasma sample from sheep (Table 1).

4.2. Test case 2: Analysis of metabolite occurrence

A laboratory which maintains a GC–MS based metabolite profiling facility will continually need to identify metabolites. Frequently, the identity of previously non-identified MSTs will be discovered and the question will arise if these MSTs were found in previous experiments [7,16,22,23] or by other laboratories [8,9]. It will therefore be important to identify the type and source of sample, which showed this MST. For this purpose, non-supervised mass spectral libraries, which may hold independently repeated analyses of each type of sample, will be valuable tools. We chose chlorogenic acid, a typical secondary product of *solanaceous* species, and the ubiquitous precursor quinic acid to demonstrate the possible gain of knowledge to be retrieved from non-supervised mass spectral libraries (Table 2). Our analysis indicated the presence above detection limit of quinic acid in almost all profiles analysed, whereas caffeic acid, the second precursor of chlorogenic acid, was present above detection limit only in leafs and *L. japonicus* nodules. In agreement with expectations, chlorogenic acid and its positional isomers were found with good mass spectral match and RI deviation in *Solanum* samples.

4.3. Test case 3: GC–MS system transfer of metabolite identifications

Almost all metabolites were analysed either in different laboratories or on two GC–MS technology platforms, GC-QUAD-MS and GC-TOF-MS. The resulting information on retention time indices from both technology platforms clearly demonstrated strict linearity in a comparative analysis of both systems, provided the same type of capillary column was used (Fig. 2). Thus, RI prediction through regression appears highly feasible for different GC–MS systems, but only when identical column types are used. Nevertheless, we detected compound specific deviations from the prediction (Fig. 2). On average we observed an error of ~5.4 RI units, but most deviations were minor and within the expected range taking the typical reproducibility of retention time indices within one system into consideration [16], namely up to 2.0 RI units (standard deviation), depending mostly on changes in metabolite amount. In

Table 2
Occurrence of metabolites in non-supervised libraries of *A. thaliana*, *L. japonicus* and *Solanum* species

Species	Organ	Quinic acid		Caffeic acid		Chlorogenic acid		4-Caffeoylquinic acid		5-Caffeoylquinic acid	
		Match	ΔRI	Match	ΔRI	Match	ΔRI	Match	ΔRI	Match	ΔRI
Arabidopsis thaliana (L.) Heynh.	Leaf	649	1.9								
Arabidopsis thaliana (L.) Heynh.	Root	723	2.0								
Lotus japonicus	Root lateral										
Lotus japonicus	Root primary										
Lotus japonicus	Nodule			752	−2.4						
Lotus japonicus	Flower	795	0.4								
Lotus japonicus	Leaf developing	685	0.1								
Lotus japonicus	Leaf mature	565[a]	0.2								
Solanum lycopersicum	Root	963	−0.2			974	−0.4				
Solanum lycopersicum	Leaf	854	0.5	838	−0.8	975	0.1	939	1.1	822	0.7
Solanum lycopersicum	Green fruit	967	1.4			970	1.6				
Solanum lycopersicum	Orange fruit	930	1.3			965	1.6				
Solanum lycopersicum	Red fruit	964	0.7			949	1.8	916	2.0		
Solanum neorickii	Fruit 45DAF	964	0.9			976	1.1	555[a]	2.1		
Solanum neorickii	Leaf	950	−0.9	862	−1.9	974	−1.2	971	0.9	827	1.0
Solanum habrochaites	Fruit 45DAF	959	0.6			949	−0.4	920	1.2		
Solanum habrochaites	Leaf	909	0.5	825	−1.9	936	0.6	962	0.9	747	0.9
Solanum parviflorum	Fruit 45DAF	964	0.8			888	0.8				
Solanum parviflorum	Leaf	926	−1.0	856	−1.6	975	−0.2	971	1.2	802	1.8
Solanum pennellii	Fruit 45DAF	799	0.0			974	−0.6				
Solanum pennellii	Leaf	953	−0.4	848	−0.6	973	1.3	925	1.5		
Solanum pimpinellifollium	Fruit 45DAF	778	−0.7			977	0.9				
Solanum pimpinellifollium	Leaf	912	−1.6	847	−2.4	966	0.1	961	1.0	840	0.4

Presence of a metabolite is validated by best mass spectral match on a scale 0–1000 (match) and smallest deviation of retention time index (ΔRI).
[a]Low mass spectral match results from mixed mass spectra with a co-eluting compound (presence of compound was manually verified).

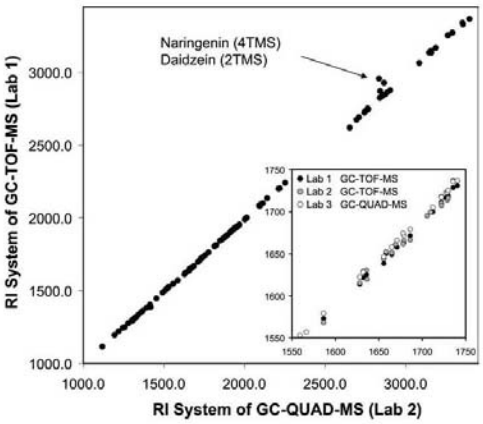


Fig. 2. Comparative analysis of retention time indices (RI), which were determined in parallel on two GC MS technology platforms in different laboratories, demonstrates that knowledge of one RI system allows good prediction of RI in a second system. Retention time indices were determined on TOF and QUAD GC MS systems. GC capillary columns were of identical build. Lab 2 operates TOF and QUAD GC MS systems.

addition, typical metabolite classes, such as sugars, fatty acids or amino acids [6–9,15,16], mostly exhibited common positive or negative trends of deviation. Therefore, RI information obtained from one technology platform will allow good prediction of retention time indices, if reference compounds are already mapped on both systems. Use and implementation of RI systems for different GC column types is ongoing effort in our laboratories (data not shown) but RI prediction will require other methods than regression, because the elution-sequence of compounds is known to change.

5. Conclusions

The hypothesis and test cases described here present, for the first time, a comprehensive MSRI library database covering MSTs of GC–MS metabolite profiles from mammals, corynebacteriae and major plant species. It includes in total more than 2000 fully evaluated mass spectral data sets obtained using two distinct technology platforms with 1089 non-redundant and 360 identified MSTs. The database is designed to be continuously extended by additional accessory information as it becomes available. We demonstrated the use of this MSRI library to screen biological samples for known compounds and showed the appliance of the non-supervised library for screening samples for known or recently identified mass spectra.

This library is constantly being updated with every new biological sample and application run in-house. Because even slight changes in GC–MS settings, such as carrier flow, temperature ramp, and dimension/make of capillary columns, induce shifts in retention behaviour of substances, GC–MS systems need to be recalibrated after each change. As there is currently no solution – other than recalibration – addressing the problem of RI shifts using different GC–MS machines we would like to offer to the biological and metabolite profiling community to perform qualitative analysis of any biological sample using our currently running protocols.

In addition to offering this service to the community, we believe that the data presented here demonstrate three general applications of such libraries, which will help to advance the field. (i) The composition of still non-characterized biological samples, for example blood plasma, or microbial extracts (data not shown) can be screened for identified constituents, and tentative best matching compounds. (ii) Occurrence of identified metabolites can be analysed in a large range of biological samples, such as different plant organs or species. For this purpose, we provide libraries comprising samples from tomato, related wild type species, and other *Solanacea*, collections of different organs of *L. japonicus*, *A. thaliana*, and preparations from microbial species. (iii) Subsequent analysis of samples on two different GC–MS systems facilitates transfer of identifications made on the first system to the second. We present data on identifications, which were made in-parallel on QUAD GC–MS and GC-TOF-MS systems in different laboratories worldwide. This is therefore the first validation that metabolite profiling, when carried out with appropriate care, can yield comparable results between laboratories. Given the number of independent laboratories involved in this study, we believe that it offers similar reassurance as provided to the microarray community by the multi-laboratory Affimetrix microbial gene expression study. We are convinced that the effort described here will be useful on several levels. Not only will it meet a recently expressed demand within the metabolomics community [13,24], which was already apparent in earliest metabolomics applications in clinical diagnostics [25], but it will also aid laboratories entering the field of metabolomics.

Acknowledgement: We thank Jim Vale and co-workers at the IGER Research Farm at Bronydd Mawr for the kind provision of sheep blood samples.

References

[1] Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G. and Schomburg, D. (2004) BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res. 32, Database issue: D431 433.
[2] Mewes, H.W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., Warfsmann, J. and Ruepp, A. (2004) MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res. 32, Database issue: D41 44.
[3] Gollub, J., Ball, C.A., Binkley, G., Demeter, J., Finkelstein, D.B., Hebert, J.M., Hernandez-Boussard, T., Jin, H., Kaloper, M., Matese, J.C., Schroeder, M., Brown, P.O., Botstein, D. and Sherlock, G. (2003) The Stanford Microarray Database: data access and quality assessment tools. Nucleic Acids Res. 31, 94 96.
[4] Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, R., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y.X., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S.G.N., Sander, C., Bork, P., Zhu, W.M., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, L., Eisenberg, D., Steipe, B., Hogue, C. and Apweiler, R. (2004) The HUPOPSI's Molecular Interaction format a community standard for the representation of protein interaction data. Nat. Biotechnol. 22, 177 183.
[5] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) The KEGG resource for deciphering the genome. Nucleic Acid Res. 32, D277 D280.
[6] Roessner, U., Wagner, C., Kopka, J., Trethewey, R.N. and Willmitzer, L. (2000) Simultaneous analysis of metabolites in

potato tuber by gas chromatography mass spectrometry. Plant J. 23, 131 142.
[7] Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L. and Fernie, A.R. (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. Plant Cell 13, 11 29.
[8] Gullberg, J., Jonsson, P., Nordstrom, A., Sjostrom, M. and Moritz, T. (2004) Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of *Arabidopsis thaliana* samples in metabolomic studies with gas chromatography/mass spectrometry. Anal. Biochem. 331, 283 295.
[9] Strelkov, S., von Elstermann, M. and Schomburg, D. (2004) Comprehensive analysis of metabolites in *Corynebacterium glutamicum* by gas chromatography/mass spectrometry. Biol. Chem. 385, 853 861.
[10] Fischer, E. and Sauer, U. (2003) Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using GC MS. Eur. J. Biochem. 270, 880 891.
[11] Sauer, U. (2004) High-throughput phenomics: experimental methods for mapping fluxomes. Curr. Opin. Biotechnol. 15, 58 63.
[12] Kromer, J.O., Sorgenfrei, O., Klopprogge, K., Heinzle, E. and Wittmann, C. (2004) In-depth profiling of lysine-producing *Corynebacterium glutamicum* by combined analysis of the transcriptome, metabolome, and fluxome. J. Bacteriol. 186, 1769 1784.
[13] Bino, R.J., Hall, R.D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B.J., Mendes, P., Roessner-Tunali, U., Beale, M.H., Trethewey, R.N., Lange, B.M., Wurtele, E.S. and Sumner, L.W. (2004) Potential of metabolomics as a functional genomics tool. Trends Plant Sci. 9, 418 425.
[14] Jenkins, H., Hardy, N., Beckmann, M., Draper, J., Smith, A.R., Taylor, J., Fiehn, O., Goodacre, R., Bino, R.J., Hall, R., Kopka, J., Lane, G.A., Lange, B.M., Liu, J.R., Mendes, P., Nikolau, B.J., Oliver, S.G., Paton, N.W., Rhee, S., Roessner-Tunali, U., Saito, K., Smedsgaard, J., Sumner, L.W., Wang, T., Walsh, S., Wurtele, E.S. and Kell, D.B. (2004) A proposed framework for the description of plant metabolomics experiments and their results. Nat. Biotechnol. 22, 1601 1606.
[15] Fiehn, O., Kopka, J., Doermann, P., Altmann, T., Trethewey, R.N. and Willmitzer, L. (2000) Metabolite profiling for plant functional genomics. Nat. Biotechnol. 18, 1157 1161.
[16] Wagner, C., Sefkow, M. and Kopka, J. (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. Phytochemistry 62, 887 900.
[17] Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmüller, E., Doermann, P., Gibon, Y., Stitt, M., Willmitzer, L., Fernie, A.R. and Steinhauser, D. (2005) GMD@CSB.DB: The Golm Metabolome Database, Bioinformatics advance access (published on December 21, 2004, doi:10.1093/bioinformatics/bti236).
[18] Ausloos, P., Clifton, C.L., Lias, S.G., Mikaya, A.I., Stein, S.E., Tchekhovskoi, D.V., Sparkman, O.D., Zaikin, V. and Zhu, D. (1999) The critical evaluation of a comprehensive mass spectral library. J. Am. Soc. Mass Spectrom. 10, 287 299.
[19] Stein, S.E. (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. J. Am. Soc. Mass Spectrom. 10, 770 781.
[20] Müller, M. and Kersten, S. (2003) Nutrigenomics: goals and strategies. Nat. Rev. Genet. 4, 315 322.
[21] Davis, C.D. and Milner, J. (2004) Frontiers in nutrigenomics, proteomics, metabolomics and cancer prevention. Mutation Res. 551, 51 64.
[22] Fiehn, O., Kopka, J., Trethewey, R.N. and Willmitzer, L. (2000) Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. Anal. Chem. 72, 3573 3580.
[23] Kopka, J., Fernie, A.R., Weckwerth, W., Gibon, Y. and Stitt, M. (2004) Metabolite profiling in plant biology: platforms and destinations. Genome Biol. 5, 109 117.
[24] Fernie, A.R., Trethewey, R.N., Krotzky, A.J. and Willmitzer, L. (2004) Innovation metabolite profiling: from diagnostics to systems biology. Nat. Rev. Mol. Cell. Biol. 5, 763 769.
[25] Jellum, E. (1977) Profiling of human-body fluids in healthy and diseased states using gas-chromatography and mass-spectrometry, with special reference to organic-acids. J. Chromatrogr. B 143, 427 462.

BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 8 2005, pages 1635–1638
doi:10.1093/bioinformatics/bti236

Systems biology

GMD@CSB.DB: the Golm Metabolome Database

Joachim Kopka[1], Nicolas Schauer[1], Stephan Krueger[1], Claudia Birkemeyer[1], Björn Usadel[1], Eveline Bergmüller[2], Peter Dörmann[1], Wolfram Weckwerth[1], Yves Gibon[1], Mark Stitt[1], Lothar Willmitzer[1], Alisdair R. Fernie[1] and Dirk Steinhauser[1,*]

[1]Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm, Germany and [2]Institute of Plant Sciences, Swiss Federal Institute of Technology, 8092, Zurich, Switzerland

Received on October 20, 2004; revised on November 16, 2004; accepted on December 15, 2004
Advance Access publication December 21, 2004

ABSTRACT

Summary: Metabolomics, in particular gas chromatography–mass spectrometry (GC–MS) based metabolite profiling of biological extracts, is rapidly becoming one of the cornerstones of functional genomics and systems biology. Metabolite profiling has profound applications in discovering the mode of action of drugs or herbicides, and in unravelling the effect of altered gene expression on metabolism and organism performance in biotechnological applications. As such the technology needs to be available to many laboratories. For this, an open exchange of information is required, like that already achieved for transcript and protein data. One of the key-steps in metabolite profiling is the unambiguous identification of metabolites in highly complex metabolite preparations from biological samples. Collections of mass spectra, which comprise frequently observed metabolites of either known or unknown exact chemical structure, represent the most effective means to pool the identification efforts currently performed in many laboratories around the world. Here we present GMD, The Golm Metabolome Database, an open access metabolome database, which should enable these processes. GMD provides public access to custom mass spectral libraries, metabolite profiling experiments as well as additional information and tools, e.g. with regard to methods, spectral information or compounds. The main goal will be the representation of an exchange platform for experimental research activities and bioinformatics to develop and improve metabolomics by multidisciplinary cooperation.

Availability: http://csbdb.mpimp-golm.mpg.de/gmd.html

Contact: Steinhauser@mpimp-golm.mpg.de

Supplementary information: http://csbdb.mpimp-golm.mpg.de/

INTRODUCTION

The sequencing and annotation of whole genomes of various organisms (Goffeau *et al.*, 1996; Blattner *et al.*, 1997; The *Arabidopsis* Genome Initiative, 2000; Lander *et al.*, 2001) facilitate the development of technology platforms to monitor the cellular inventory (Fiehn *et al.*, 2000; Lockhart and Winzeler, 2000; Corbin *et al.*, 2003). Since the dawn of genomic technology in the past decade and in conjunction with enhancing genomic information a vast amount of diverse data has been generated and released to the public community. The improving knowledge of gene functions in concurrence with global expression analyses allows phenotypes to be linked to their co-responding genomic data. However, our knowledge of the molecular basis of biological functions and their respective contribution to observed phenotypes is, as yet, relatively rudimentary. The recent mining and exploitation of data by multiparallel 'omics' technologies open up the possibility to gain comprehensive insight into the understanding of biological systems (Kitano, 2002; Oltvai and Barabási, 2002; Fernie *et al.*, 2004). The flood of information obtained worldwide by scientists for this purpose urgently requires user-friendly public data access. In the past decades much progress has been made on the storage of information derived from the various levels of the cellular hierarchy. For instance, databases like BRENDA (Schomburg *et al.*, 2004), KEGG (Kanehisa *et al.*, 2004) or MetaCyc (Krieger *et al.*, 2004) harbour information concerning metabolic pathways, chemical reactions including inventories of the genes and enzymes involved. Genomic databases, such as MIPS (Mewes *et al.*, 2004), TAIR (Rhee *et al.*, 2003) and TIGR (Quackenbush *et al.*, 2000), provide public access to protein sequences based on whole genome analyses, maps of protein–protein interactions, protein localization and many further features. Developments in transcript profiling technologies have led to the adoption of uniform experimental platforms that are used worldwide. The widely shared experimental approach facilitated the establishment of expression profile related databases, such as the Stanford Microarray Database [SMD (Gollub *et al.*, 2003)], TAIR or NCBI–GEO (Edgar *et al.*, 2002). Similarly the availability of proteome data has led to the establishment of various databases [e.g. Swiss-Prot (Boeckmann *et al.*, 2003)] or initiatives [e.g. HPI, (Hermjakob *et al.*, 2004)] which focus on the functional annotation of proteins and standardization of protein data.

In contrast to the multitude of well established databases which comprise information on the genome, transcriptome and proteome, no attempt has been made to store the flood of data arising from metabolome analyses of biological samples. As already outlined, metabolites have an enormous diversity of chemical structures. These are identified and quantified using a wide range of technology platforms (Kopka *et al.*, 2004). Thus there is an urgent need for publicly accessible metabolome databases that harbour underlying information. Here we describe the Golm Metabolome Database (GMD), an

*To whom correspondence should be addressed.

open access database for exchange and presentation of metabolomic and related information. In the current build GMD focuses on gas chromatography–mass spectrometry (GC–MS) (Roessner *et al.*, 2000), the most advanced and widespread technology platform for metabolomics. The collected information (1) covers knowledge concerning analytical technologies and (2) harbours information that supports unequivocal metabolite identification. In addition, GMD provides access to stored metabolite profiles.

SYSTEMS OVERVIEW

Affiliation and implementation

The GMD platform is affiliated to CSB.DB, a comprehensive systems-biology database, which is hosted at the Max-Planck-Institute of Plant Molecular Physiology, Potsdam–Golm, Germany (Steinhauser *et al.*, 2004). GMD complements the currently available transcriptional co-response databases and uses a similar system for data storage and handling.

Information on analytical technologies

The highly complex nature and the enormous chemical diversity of compounds obtained when analyzing the metabolome of organisms constitutes one of the main challenges in metabolomics (Oksman-Caldentey *et al.*, 2004; Fernie *et al.*, 2004). Current estimations vary. However, 4000–25 000 compounds may represent the metabolome of any given organism (Trethewey, 2004; Fernie *et al.*, 2004). The plant kingdom is believed to comprise in excess of 200 000 metabolites (Fiehn, 2002; Trethewey, 2004). Highly diverse chemical characteristics in conjunction with the vast amount of potential compounds have profound implications for metabolite extraction and stability. Any given protocol for metabolome measurement thus represents a well tuned balance between accuracy and metabolite coverage. The GMD analytics pages allow access to expert knowledge on methods applied by the GMD contributors. Information on different technology platforms, publicly available methods, as well as contact information for individual knowledge exchange is included. Furthermore, an overview of the available resources is given for those scientists who intend to enter the field of experimental physiology and plan to set up a metabolomics facility.

Mass spectra and retention time index (MSRI) libraries

Following analytical measurements, data processing algorithms are applied to detect metabolic components in spectral data. The identification and characterization of the hundreds to thousands of metabolites obtained from diverse biological samples represents a major challenge in metabolomics. These identification efforts require large-scale processing of pure standard substances to generate customized spectral libraries that can be used for subsequent identification of hitherto unknown metabolic components from spectral data. To overcome the current limitation of customized mass spectral libraries that need to be maintained by each laboratory the GMD mass spectra information pages are developed to exchange information. In detail, we started to disseminate the underlying evidences that support metabolite identification in complex GC–MS profiles from diverse biological sources. The MSRI web platform provides access to customized MSRI libraries, which were generated using identical capillary GC columns and settings using two different electron impact ionization GC–MS technologies, namely quadrupole GC–MS (Fiehn *et al.*, 2000; Roessner *et al.*, 2000) and GC–TOF (time-of-flight)–MS (Wagner *et al.*, 2003; Weckwerth *et al.*, 2004). Currently, five downloadable libraries are available, which may be imported into the NIST02 mass spectral search program or AMDIS, a technology platform independent automated mass spectral deconvolution and identification system (National Institute of Standards and Technology, Gaithersburg, MD, USA). The above libraries are split according to the technology platform and the degree of manual mass spectral curation. The Q_MSRI and T_MSRI libraries contain mass spectral tags (MSTs), which were either generated on four identically configured quadrupole GC–MS systems (Q_MSRI) or on a single time-of-flight system (T_MSRI). Mass spectral libraries, which exclusively consist of manually evaluated, identified or classified MSTs are assigned to ID-libraries. In contrast, libraries which were generated by automated deconvolution were assigned to NS libraries, indicative of the non-curated mode of construction. The currently available libraries cover data from mammals, yeast, corynebacterium, model plants, such as crop plants and related wild species, as well as required non-sample controls. These libraries currently feature more than 2000 evaluated mass spectra from the two technology platforms which represent 1089 non-redundant and 360 identified MSTs.

The metabolite profiling platform (GMD profiles)

The vast amount of complex data obtained from metabolite profiling experiments in conjunction with the ongoing developments on analytical technologies require the public availability of these data for cross-comparison and cross-experiment analysis. According to these demands we started to present metabolic fingerprinting and metabolite profiling experiments, which can be currently searched by compound names or browsed by a list of experiments.

For the exchange of the highly complex experimental background information and data from metabolite profiling experiments we implemented the MIAMET description as suggested by Bino *et al.* (2004). For future implementations and development of the GMD platform recently made advances in database modelling and insight into the architecture of metabolomics data (ArMet) will prove to be highly important (Jenkins *et al.*, 2004).

IMPLEMENTATION AND QUERY OVERVIEW

Content browsing and queries

The GMD content can be explored by browsing the HTML content through lists or a simple site map, represented as a hierarchical tree, which is linked to the available second level of HTML pages. Information regarding downloadable MSRI libraries as well as related Supplementary information, such as technologies, method descriptions and acknowledgements, are made accessible. Both, the MSRI libraries as well as the currently integrated metabolite profiling experiments are presented in table format, which provides links to associated detailed information.

A more sophisticated way to explore the GMD content is offered through the available query pages. Currently, five different types of queries are implemented which can all be accessed by the GMD site map.

MSRI compound search

The compound search tool allows searching by compound name and provides access to the linked mass spectral information harboured at GMD. Various filter options can be applied to restrict the query results, e.g. to the available technology platforms, particular libraries or methods. The retrieved mass spectral entries are presented as a table which contains basic mass spectral information for a particular compound, such as compound role, i.e. metabolite or internal standard, observed retention time index (RI) and technology platform. This basic information can be sorted upon user invocation. All information is linked to the detailed physicochemical characteristics of the available mass spectra, which are represented as a mass spectrum chart. This final level of information facilitates the identification of compounds in profile analyses. The in-depth mass spectral information encompasses in addition (1) the recommended quantifier and qualifier masses, (2) access to available replicate mass spectra of the same compound and (3) a direct link to the mass spectrum search and comparison tool.

MSRI mass spectrum search

For analysis of mass spectra that are present in the libraries but can also be user-submitted we implemented a query tool which allows comparison with all available curated mass spectra of our libraries. Mass spectra may be submitted in either NIST02 or AMDIS format (Ausloos *et al*., 1999; Stein, 1999). The search is performed by computing the fragment-intensity agreement, measured as dynamically normalized Euclidean distance [Euclid], as S12 [s12] index (Gower and Legendre, 1986), Hamming (1950) and Jaccard (1908) distance. The result set is presented as a sortable HTML table containing information such as the rank, the identifier for each spectrum, the RI, the method information, the compound name in case of identified metabolites and all computed similarity measures. All types of information can be used for sorting. Moreover, additional criteria for comparison are given based on absolute RI differences to (1) the observed RI as provided by an optional user input and (2) as calculated relative to the best hit. If available, occurrence of qualifier as well as quantifier masses is considered. A head-to-tail plot of the query and selected hit spectra can be invoked. Depending on the chosen sorting a colour-coded graphical representation of the ten best hits is generated below the result table. The graphical output is similar to a typical BLAST (Altschul *et al*., 1990) result. The ratio plot mirrors the occurrence of the masses and their co-responding ratios of intensities in comparison to the query spectrum. The result table can be downloaded by an exporter function as a tab-delimited and zip-coded file. The file contains all data presented in HTML table and in addition all returned mass spectra of the query. Various filter options, especially restriction to a predefined RI window or set of major fragments, can be invoked by the user to limit the search to relevant results. The set of implemented tools is complementary to those available within the NIST02 software.

MSRI customized library generation

In extension to the precompiled MSRI download libraries GMD allows the generation of user-customized mass spectral subsets from the full repository of curated mass spectral entries. These subsets can be downloaded as a zip-coded text file and treated like our precompiled MSRI libraries (see above). The search input is currently restricted to MPIMP-Ids, which can be obtained through the above-mentioned queries or by using the compound name converter (see below). We suggest limiting of results according to the GC–MS technology platform or the analytical methods used in order to obtain the curated spectra.

Profile compound search

As mentioned above GMD has started to integrate a first set of metabolite profiling experiments which were generated with a quadrupole GC–MS technology platform. Currently, 69 profiles of nine replicate sets are included describing metabolic changes under different light conditions. The profiles can be queried by compound name and allows searches for the changes in compound levels. Various filter options are available to restrict computation to high-quality mass traces by using the default or user modified values. Moreover, the user can select between parametric or non-parametric statistics for the dynamic computation of the treatment versus control comparisons. The result set covers information on experimental background, performed comparisons as well as information on significance of the observed differences. Furthermore, treatment versus control ratios are given and colour coded to mark decreases or increases. In analogy to the Affymetrix oligonucleotide technology platform we use different masses as representatives for any particular compound. All used masses are represented in the result tables and are checked for co-responding behaviour across the full experimental dataset. Future updates will connect metabolite profiles of GMD to the visualization software tools MapMan (Thimm *et al*., 2004) and PaVESy (Ludemann *et al*., 2004).

Compound name converter

Because of the different usage of compound names and identifiers in the publicly available databases we implemented a converter which allows converting of user compound names to MapMan names for a stand-alone visualization of the results with the MapMan software (Thimm *et al*., 2004) or to MPIMP-Ids for customized library generation.

OUTLOOK

GMD will frequently be updated with new mass spectra, metabolite identifications, mass spectral libraries of biological samples and metabolite profiling experiments. GMD is intended as a repository for experiments performed at the Max-Planck-Institute of Molecular Plant Physiology and for data made available through collaborating scientists. We offer our already well-characterized GC–MS technology platforms specifically for cooperations on metabolite identification in complex biological samples. As suggested by Bino *et al*. (2004) we envision to share biological samples and metabolite identifications between laboratories engaged in GC–MS metabolite profiling. Thus we provide a public platform for future advances and developments in metabolomic science. In-depth analysis and understanding of metabolome data at systems level will require a multidisciplinary effort, especially integration of proteome and transcriptome data. Such interdisciplinary cooperation and data mining is in preparation and in the case of steady state transcript analysis already in place (Steinhauser *et al*., 2004). We are convinced that GMD will represent a crucial building block for CSB.DB (http://csbdb.mpimp-golm.mpg.de). CSB.DB, a comprehensive systems-biology database project, will harbour and allow joined access to metabolome, proteome and transcriptome data. Thus CSB.DB will develop into a highly useful and informative public

resource for researchers focusing on experimental biology as well as for computational biology and bioinformatics.

ACKNOWLEDGEMENTS

We appreciate the work of all scientists, who contributed samples and submitted mass spectral information or metabolite profiling experiments to GMD. Detailed acknowledgements and affiliations are made accessible through GMD (http://csbdb.mpimp-golm.mpg.de/gmd.html). We are grateful to the Max-Planck-Institute of Molecular Plant Physiology and the Max-Planck-Society for long-standing and continuous support of the Golm Metabolome Database.

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignments search tool. *J. Mol. Biol.*, **215**, 403–410.

Ausloos,P., Clifton,C.L., Lias,S.G., Mikuya,A.I., Stein,S.E., Tchekhovskoi,D.V., Sparkman,O.D., Zaikin,V. and Zhu,D. (1999) The critical evaluation of a comprehensive mass spectral library. *J. Am. Soc. Mass Spectrom.*, **10**, 287–299.

Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.

Bino,R.J., Hall,R.D., Fiehn,O., Kopka,J., Saito,K., Draper,J., Nikolau,B.J., Mendes,P., Roessner-Tunali,U., Beale,M.H. *et al.* (2004) Potential of metabolomics as a functional genomics tool. *Trend Plant Sci.*, **9**, 418–425.

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Corbin,R.W., Paliy,O., Yang,F., Shabanowitz,J., Platt,M., Lyons,C.E.,Jr, Root,K., McAuliffe,J., Jordan,M.I., Kustu,S. *et al.* (2003) Toward a protein profile of *Escherichia coli*: comparison to its transcription profile. *Proc. Natl Acad. Sci. USA*, **100**, 9232–9237.

Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Fiehn,O., Kopka,J., Dormann,P., Altmann,T., Trethewey,R.N. and Willmitzer,L. (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.*, **18**, 1157–1161.

Fiehn,O. (2002) Metabolomics—the link between genotypes and phenotypes. *Plant Mol. Biol.*, **48**, 155–171.

Fernie,A.R., Trethewey,R.N., Krotzky,A.J. and Willmitzer,L. (2004) Metabolite profiling: from diagnostics to systems biology. *Nat. Rev. Mol. Cell Biol.*, **5**, 763–769.

Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1996) Life with 6000 Genes. *Science*, **274**, 546–567.

Gollub,J., Ball,C.A., Binkley,G., Demeter,J., Finkelstein,D.B., Hebert,J.M., Hernandez-Boussard,T., Jin,H., Kaloper,M., Matese,J.C. *et al.* (2003) The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.*, **31**, 94–96.

Gower,J.C. and Legendre,P. (1986) Metric and Euclidean properties of dissimilarity coefficients. *J. Classif.*, **3**, 5–48.

Hamming,R.W. (1950) Error detecting and error correcting codes. *Bell Syst. Tech. J.*, **9**, 147–160.

Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *et al.* (2004) The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol.*, **22**, 177–183.

Jaccard,P. (1908) Nouvelles recherches sur la distribution florale. *Bull Soc. Vaud Sci. Nat.*, **44**, 223–270.

Jenkins,H., Hardy,N., Beckmann,M., Draper,J., Smith,A.R., Taylor,J., Fiehn,O., Goodacre,R., Bino,R., Hall,R. *et al.* (2004) A proposed framework for the description of plant metabolomics experiments and their results. *Nat. Biotechnol.*, **22**, 1601–1606.

Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.

Kitano,H. (2002) Systems biology: a brief overview. *Science*, **295**, 1662–1664.

Kopka,J., Fernie,A., Weckwerth,W., Gibon,Y. and Stitt,M. (2004) Metabolite profiling in plant biology: platforms and destinations. *Genome Biol.*, **5**, 109.

Krieger,C.J., Zhang,P., Mueller,L.A., Wang,A., Paley,S., Arnaud,M., Pick,J., Rhee,S.Y. and Karp,P.D. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **32**, D438–D442.

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., Fitzhugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lockhart,D.J. and Winzeler,E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–836.

Lüdemann,D., Weicht,D., Selbig,J. and Kopka,J. (2004) PaVESy: pathway visualization and editing system. *Bioinformatics*, **20**, 2841–2844.

Mewes,H.W., Amid,C., Arnold,R., Frishman,D., Guldener,U., Mannhaupt,G., Munsterkotter,M., Pagel,P., Strack,N., Stumpflen,V. *et al.* (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, **32**, D41–D44.

Oksman-Caldentey,K.-M., Inzé,D. and Oresič,M. (2004) Connecting genes to metabolites by a systems biology approach. *Proc. Natl Acad. Sci. USA*, **101**, 9949–9950.

Oltvai,Z.N. and Barabási,A.-L. (2002) Life's Complexity Pyramid. *Science*, **298**, 763–764.

Quackenbush,J., Liang,F., Holt,I., Pertea,G. and Upton,J. (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.*, **28**, 141–145.

Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. *et al.* (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.

Roessner,U., Wagner,C., Kopka,J., Trethewey,R.N. and Willmitzer,L. (2000) Simultaneous analysis of metabolites in potato tuber by gas chromatography–mass spectrometry. *Plant J.*, **23**, 131–142.

Schomburg,I., Chang,A., Ebeling,C., Gremse,M., Heldt,C., Huhn,G. and Schomburg,D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.*, **32**, D431–D433.

Stein,S.E. (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/ mass spectrometry data. *J. Am. Soc. Mass Spectrom.*, **10**, 770–781.

Steinhauser,D., Usadel,B., Luedemann,A., Thimm,O. and Kopka,J. (2004) CSB.DB: a comprehensive systems-biology database. *Bioinformatics*, **20**, 3647–3651.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Thimm,O., Blasing,O., Gibon,Y., Nagel,A., Meyer,S., Kruger,P., Selbig,J., Muller,L.A., Rhee,S.V. and Stitt,M. (2004) MAPMAN: a user-driven tool to display genomics datasets onto diagrams of metabolic pathways and other biological processes. *Plant J.*, **37**, 914–939.

Trethewey,R.N. (2004) Metabolite profiling as an aid to metabolic engineering in plants. *Curr. Opin. Plant Biol.*, **7**, 196–201.

Wagner,C., Sefkow,M. and Kopka,J. (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI–TOF–MS metabolite profiles. *Phytochemistry*, **62**, 887–900.

Weckwerth,W., Loureiro,M.E., Wenzel,K. and Fiehn,O. (2004) Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl Acad. Sci. USA*, **101**, 7809–7814.

BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 20 2005, pages 3887–3895
doi:10.1093/bioinformatics/bti634

Gene expression

Non-linear PCA: a missing data approach

Matthias Scholz[1,*], Fatma Kaplan[2], Charles L. Guy[2], Joachim Kopka[1] and Joachim Selbig[1,3]

[1]Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany, [2]University of Florida, Plant Molecular and Cellular Biology Program, Department of Environmental Horticulture, Gainesville, Florida 32611, USA and [3]University of Potsdam, Bioinformatics, Germany

Received on January 5, 2005; revised on August 2, 2005; accepted on August 15, 2005
Advance Access publication August 18, 2005

ABSTRACT

Motivation: Visualizing and analysing the potential non-linear structure of a dataset is becoming an important task in molecular biology. This is even more challenging when the data have missing values.

Results: Here, we propose an inverse model that performs non-linear principal component analysis (NLPCA) from incomplete datasets. Missing values are ignored while optimizing the model, but can be estimated afterwards. Results are shown for both artificial and experimental datasets. In contrast to linear methods, non-linear methods were able to give better missing value estimations for non-linear structured data.

Application: We applied this technique to a time course of metabolite data from a cold stress experiment on the model plant *Arabidopsis thaliana*, and could approximate the mapping function from any time point to the metabolite responses. Thus, the inverse NLPCA provides greatly improved information for better understanding the complex response to cold stress.

Contact: scholz@mpimp-golm.mpg.de

1 INTRODUCTION

Non-linear principal component analysis (NLPCA) is generally seen as a non-linear generalization of standard linear principal component analysis (PCA) (Jolliffe, 1986; Diamantaras and Kung, 1996). The principal components are generalized from straight lines to curves. Here, we focus on a neural network based NLPCA, the auto-associative neural network (Kramer, 1991; DeMers and Cottrell, 1993; Hecht-Nielsen, 1995; Kirby and Miranda, 1996; Malthouse, 1998). It is successfully applied in the fields of atmospheric and oceanic sciences (Hsieh, 2004; Monahan *et al.*, 2003), in astronomy and even in biomedical research. In Scholz and Vigário (2002) a hierarchically extended version of NLPCA was applied to spectral data from stars and to electromyographic (EMG) recordings for different muscle activities.

There is a wide variety of methods for visualizing data and extracting meaningful components (also termed features, factors or sources) in a non-linear way. Locally linear embedding (LLE) (Roweis and Saul, 2000; Saul and Roweis, 2004) and Isomap (Tenenbaum *et al.*, 2000) were developed to visualize high dimensional data by projecting (embedding) them into a two- or low-dimensional space. A mapping function as a non-linear model is not explicitly given. Principal curves (Hastie and Stuetzle, 1989) and self-organizing maps (SOM) (Kohonen, 2001) are useful for detecting non-linear curves and two-dimensional non-linear planes. Both methods are limited to extraction of two components at most, due to high computational costs. Kernel PCA (Schölkopf *et al.*, 1998), when used as pre-processing, can improve classification results.

Here, we consider the neural network approach. It provides a non-linear model of the mapping function and we will show that it can be applied to incomplete datasets by modelling only the second part of the auto-associative network, the reconstruction or generation part. The difficulty is to estimate both the model weights and the inputs which are now the required components.

For this approach Hassoun and Sudjianto (1997) optimized the weights and the inputs in two alternate steps by minimization of an error function which is equivalent to maximum likelihood. A similar approach was also used by Oh and Seung (1998). As the inputs can be represented by weights, we propose to optimize the inputs and weights simultaneously.

The same network architecture is also used by Valpola for a non-linear factor analysis (NFA) and a non-linear independent factor analysis (NIFA) (Lappalainen and Honkela, 2000; Honkela and Valpola, 2005), also applicable to incomplete datasets (Raiko and Valpola, 2001). The weights and inputs are optimized by Bayesian learning. The inputs (components) are explicitly modelled by a plain Gaussian distribution in NFA and a mixture of Gaussian distribution in NIFA. Although Bayesian inference in NFA and maximum likelihood in NLPCA often lead to similar results, their conceptual basis is rather different. Maximum likelihood attempts to find a single set of values for the network weights and inputs. In contrast, in the Bayesian approach the weights and inputs are described by posterior probability distributions which lead to a good regularisation. There are some relations: the Gaussian prior distribution for the weights corresponds to the use of a weight-decay regularizer in the maximum likelihood approach. Minimization of a mean square error function is equivalent to one maximum a posteriori (MAP) with additive Gaussian observation noise. In the proposed inverse

*To whom correspondence should be addressed.

NLPCA model a single error function is minimized. The model weights and inputs (components) are optimized simultaneously and the model is extended to be applicable to incomplete datasets. There are many methods for estimating missing values (Little and Rubin, 2002). Here, we focus on detecting non-linear components from incomplete datasets, so our approach involves ignoring missing values not a priori estimating them. However, once the non-linear mapping is effectively modelled, the missing values can then be estimated as well. This is shown for an artificial dataset and for experimental data. Estimation results were compared with results of state-of-the-art estimation techniques. There are two PCA based linear techniques: the recently published Bayesian missing value estimation method for gene expressions (Oba *et al.*, 2003) which is based on Bayesian principal component analysis (BPCA) (Bishop, 1999) and probabilistic PCA (PPCA) (Verbeek *et al.*, 2002) based on Roweis, (1997). Furthermore, there are the *k*-nearest neighbour based approach, KNNimpute (Troyanskaya *et al.*, 2001), and a non-linear estimation by SOM.

There are many other approaches which are not further considered; for example, there are methods based on non-linear regression among variables (Zhou *et al.*, 2003) or on modelling a dynamical system (Simeka and Kimmel, 2003). The latter takes the time information into account. It belongs, therefore, to supervised methods where it is much more difficult to avoid over-fitting than in the previously mentioned unsupervised methods.

Cold stress to the cell can cause rapid changes in metabolite levels. Here, we have analysed the temporal metabolite response to cold stress in the model plant *Arabidopsis thaliana*. The proposed inverse NLPCA model was applied to these, partly incomplete, metabolite data (Kaplan *et al.*, 2004). Thus, we model the cold stress adaptation by a mapping function from a given time point to the metabolite responses. For each time point we are able to give the metabolites in the order of importance, i.e. the metabolites are ranked by the relative change in their concentration level. This procedure is analogous to ranking in PCA by the eigenvector values (also termed loadings or weights).

The observed experimental time information is not used in this unsupervised model. Thus, the risk of over-fitting is much lower than in a supervised regression model. Furthermore, the response time and developmental state of plant individuals in any experiment differs from the exact physical time measurement. Hence we cannot absolutely trust the physical experimental time for the description of biological experiments. An unsupervised model will be superior in accommodating the unavoidable individual variability of biological samples such as plants.

2 AUTO-ASSOCIATIVE NEURAL NETWORKS

The NLPCA, proposed by Kramer (1991), is based on a multi-layer perceptron (MLP) with an auto-associative topology, also known as an autoencoder, replicator network, bottleneck or sand glass type network. A good introduction to multi-layer perceptrons can be found in Bishop (1995), Haykin (1998).

The auto-associative network performs the identity mapping, the output $\hat{x}$ has to be equal to the input x, by minimizing the square error $\|x-\hat{x}\|^2$.

This is no trivial task, as there is a 'bottleneck' in the middle, a layer of fewer nodes than at input or output, where the data have to be projected or compressed into a lower dimensional space Z, (Fig. 1).

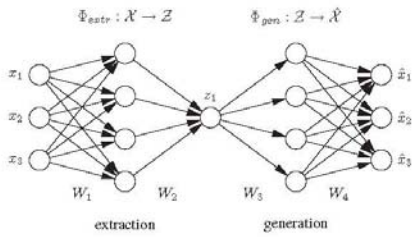


Fig. 1. The standard auto-associative neural network. The network output $\hat{x}$ is required to be equal to the input x. Illustrated is a [3-4-1-4-3] network architecture. Biases have been omitted for clarity. Three-dimensional samples x are compressed (projected) to one component z by the extraction part. The inverse generation part reconstructs $\hat{x}$ from z. The sample $\hat{x}$ is usually a noise-reduced representation of x.

The network can be divided into two parts: the first part represents the extraction function $\Phi_{extr} : \mathcal{X} \to \mathcal{Z}$, whereas the second part represents the inverse function, the generation or reconstruction function $\Phi_{gen} : \mathcal{Z} \to \hat{\mathcal{X}}$. A hidden layer in each part enables the network to perform non-linear mapping functions.

3 INVERSE NLPCA MODEL

The inverse model of NLPCA extracts the required components by only modelling the generation function $\Phi_{gen} : \mathcal{Z} \to \hat{\mathcal{X}}$ of the auto-associative network. It is the inverse function to the component extraction function $\Phi_{extr} : \mathcal{X} \to \mathcal{Z}$.

The inverse model presents a set of advantages; we only have to train the second part of the auto-associative network, which is more efficient than training both parts. Also, we model the natural process, which has generated the observed samples, hence we can be sure that such a function exists, which is not necessarily the case for the extraction model. And, most importantly, the inverse NLPCA can be extended to handle incomplete datasets, as we do not need the sample data as input, the data are needed only as required output.

As the desired components are now unknown inputs, the blind inverse problem is to estimate both the inputs and the parameters of the model by only given outputs. This makes sense only with the additional constraint of a lower dimensional input.

The output $\hat{x}$ depends on the input z and the network weights $w \in W_3$, W_4, as illustrated in Figure 2,

$$\hat{x} = \Phi_{gen}(w, z) = W_4 g(W_3 z)$$

The non-linear activation function g (e.g. tanh) is applied element-wise. Biases are not explicitly considered; however, they can be included by introducing an extra unit, or input, with activation fixed at one. The mean square error depends on z and w as well:

$$E(w, z) = \frac{1}{dN} \sum_{n}^{N} \sum_{i}^{d} \left[x_i^n - \sum_{j}^{h} w_{ij} g\left(\sum_{l}^{m} w_{jk} z_k^n \right) \right]^2,$$

d is the dimensionality of the data (the number of metabolites), N is the number of samples.

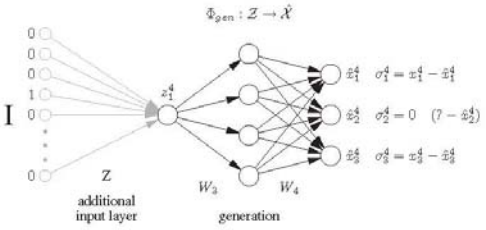


Fig. 2. The proposed inverse NLPCA model as [1-4-3] network. Only the generation part (black) of the auto-associative network (Fig. 1) is used. The inputs z can be optimized by propagating the partial errors back to the input layer z. This is equivalent to the illustrated prefixed input layer (grey), where the weights are representing the component values z. The input is now a (sample x sample) identity matrix I. For the 4th sample ($n=4$), as illustrated, all inputs are zero except the 4th, which is one. On the right, the second element x_2^4 of the 4th sample x^4 is missing. Therefore, the partial error σ_2^4 is set to zero, identical to ignoring or non-back-propagating.

The error can be minimized by a gradient optimization algorithm, e.g. conjugate gradient descent (Hestenes and Stiefel, 1952; Press *et al.*, 1992). The gradients are obtained by propagating the partial errors σ_i^n back to the input layer. For the input gradients it is simply one step further than usual. The gradients of the weights $w_{ij} \in W_4$, $w_{jk} \in W_3$ and inputs z_k^n are the partial derivatives:

$$\frac{\partial E}{\partial w_{ij}} = \sum_n \sigma_i^n g(a_j^n); \quad \sigma_i^n = \hat{x}_i n - x_i^n$$
$$\frac{\partial E}{\partial w_{jk}} = \sum_n \sigma_j^n z_k^n; \qquad \sigma_j^n = g'(a_j^n) \sum_i w_{ij} \sigma_i^n$$
$$\frac{\partial E}{\partial z_k^n} = \sigma_k^n; \qquad \sigma_k^n = \sum_j w_{kj} \sigma_j^n$$

For the bias, additional weights w_{i0} and w_{j0} can be used, with associated constants $z_0 = 1$ and $g(a_0) = 1$. The weights w and the inputs z can be optimized simultaneously, by considering (w, z) as one vector to optimize with given gradients. This would be equivalent to an approach where an additional input layer is representing the components z as weights, and new inputs are given by a (sample x sample) identity matrix, as illustrated in Figure 2. However, this layer is not needed for implementation. The purpose of the additional input layer is only to explain that the inverse NLPCA model can be converted to a conventionally trained multi-layer perceptron, with known inputs and simultaneously optimized weights, including the weights z, representing the desired components. Hence, an alternating approach as done by Hassoun and Sudjianto (1997) is unnecessary. Beside a more efficient optimization, it also avoids the risk of oscillating while training in an alternating approach.

A disadvantage of such an inverse approach is that there is no mapping function $\mathcal{X} \to \mathcal{Z}$, required for new data x. However, we can approximate the mapping by searching for an optimal input z to a given new sample x. For that, the network weights w have to be fixed and the input z has to be optimized to minimize the square error $\|x - \hat{x}(z)\|^2$. This is only a line search (in case of one component) or low dimensional optimization with given gradients, efficiently done by a gradient optimization algorithm.

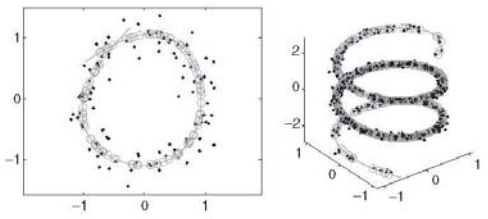


Fig. 3. Approximation of a circular (left) and a helical (right) structure by the proposed inverse NLPCA model. The noisy data x (dots) are projected onto a one-dimensional non-linear component (line). The projection or de-noised reconstruction $\hat{x}$ is marked by a circle. Note that an inverse model is able to extract self-intersecting components (left).

The inverse NLPCA is able to extract components of higher non-linear complexity than the standard NLPCA, even self-intersecting components can be modelled. This is shown in Figure 3 for a circular structure in two dimensions, generated from a uniformly distributed factor t (the angle) and a helical structure embedded in three dimensions, generated from a Gaussian distributed factor t. For the uniformly distributed 100 circular data points (plus noise), a [1-3-2] network is trained in 3000 iterations. The noisy helical structure of 1000 Gaussian distributed data points, is modelled with a [1-8-3] network in 10 000 iterations.

The inverse NLPCA is not restricted to one component. It can be extended to m components by increasing the number of units in the input layer, the component layer z, to m. With an additional hierarchical error function (Scholz and Vigário, 2002), the non-linear components $1, \ldots, m$ can be extracted in a hierarchical order, which is a natural non-linear extension to the hierarchical ordered components of the standard linear PCA.

3.1 Regularization

As we usually have a large number of dimensions (metabolites) and a relatively small number of samples, a regularization of the non-linear model is very important.

Standard methods for regularization in neural networks reduce the number of hidden units or add a weight decay term to the error function. Furthermore, auto-associative neural networks have a kind of self-regularization, caused by the fact that for each mapping function the inverse function has to be estimated as well. A complex function has usually a much more complex inverse function or the inverse function does not even exist. Therefore, the auto-associative neural network is constrained to keep the functions as simple as possible. A similar effect is observed when extracting non-linear components in a hierarchical order, where subsequent components are extracted in respect to the previous components. A complex first component would strongly increase the complexity of the second or later components. Thus, the network is constrained to generate very smooth first components.

4 MISSING VALUE ESTIMATION

The inverse NLPCA model can be easily extended to be applicable to incomplete datasets. If the ith element x_i^n of the nth sample vector x^n is missing, the partial error σ_i^n is set to zero before

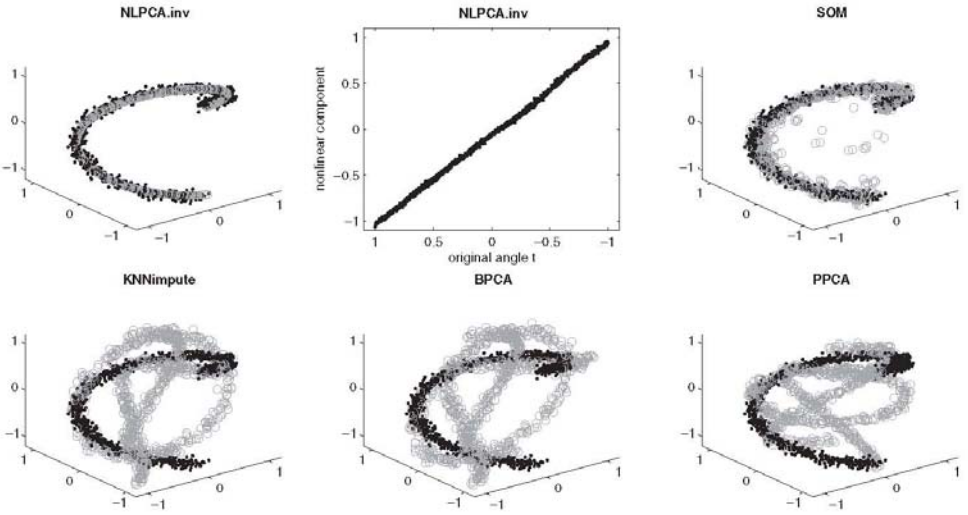


Fig. 4. Artificial data were generated to test different missing value algorithms. The samples form a helical loop. From each of the three-dimensional samples, one value is removed and then estimated by each missing value algorithm. The known complete samples are plotted as dots and the estimated values as circle. Above: the inverse NLPCA is able to extract the non-linear component from this highly incomplete dataset, and hence it can give a very good estimation of the missing values. SOM also gives a reasonably good estimation, but the linear approaches BPCA and PPCA, as well as the k-nearest neighbour based approach KNNimpute, fail with this non-linear dataset, see also Table 1.

back-propagating; hence this error is ignored, and it has no contribution to the gradients. Thus, the non-linear components are extracted from all the available observations. With these components the original data can be reconstructed, including the missing values. The network output x_i' gives the estimation of the missing element x_i^n.

4.1 Missing data: artificial data

The inverse NLPCA approach was first applied to an artificial dataset and the results were compared with other missing value estimation techniques, the linear techniques BPCA[1] and PPCA[2], the k-nearest neighbour based approach KNNimpute[3], and the non-linear SOM[4]. The data x lie on a one-dimensional manifold (a helical loop) embedded in three dimensions, plus Gaussian noise with standard deviation 0.05, see Figure 4. 1000 samples x were generated from an uniformly distributed factor t over the range $[-1, 1]$, t represents the angle:

$$x_1 = \sin(\pi t)$$
$$x_2 = \cos(\pi t)$$
$$x_3 = t.$$

From each three-dimensional sample, one value is randomly removed and is regarded as missing. This gives a high missing value rate of 33.3 percent. However, if the non-linear component (the helix) is known, the estimation of a missing value is given exactly by the two other coordinates, except at the first and last positions of the helix loop, where in the case of missing vertical coordinate x_3, the sample can be assigned either to the first or to the last position. There are two possible optimal solutions; consequently, missing value estimation is not always unique in the non-linear case.

In Figure 4 and Table 1 it is shown that even if the datasets are incomplete for all samples, the inverse NLPCA model is able to detect the non-linear component and gives a very good missing

[1] http://hawaii.aist-nara.ac.jp/~shige-o/tools/
[2] http://carol.science.uva.nl/~jverbeek/software/
[3] http://smi-web.stanford.edu/projects/helix/pubs/impute/
[4] http://www.cis.hut.fi/projects/somtoolbox/

Table 1. MSE of missing value estimation

	Noise	Noise-free
NLPCA.inv	0.0021	0.0013
SOM	0.0405	0.0384
KNNimpute	0.4435	0.4429
BPCA	0.4191	0.4186
PPCA ($k = 3$)	0.4354	0.4347
Mean	0.4429	0.4422

Mean square error (MSE) of different missing value estimation techniques, applied to the helical data (Fig. 4). The inverse NLPCA model gives a very good estimation of the missing values. Although the model was trained with noisy data, the noise-free data were better represented than the noisy data, confirming the de-noising ability of the model.

Also SOM gives a good estimation, but the linear techniques BPCA and PPCA, as well as KNNimpute are not able to give good estimations, the results are similar to the results of naive substitution by the mean over the residuals of one variable.

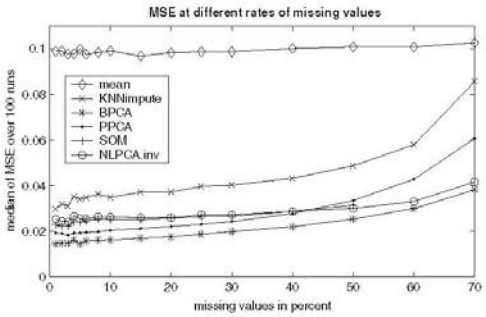


Fig. 5. From an experimental dataset of completely available 140 metabolites, different percentages of values were removed randomly and estimated by different missing value algorithms. This is done 100 times with differently removed values. The MSE over all the runs is plotted. An estimation by mean over the residual values gives the worst result. It is used as a base line. BPCA gives the best result. However, this is the case only when all 140 metabolites are considered including the large number of non-relevant metabolites with small relative variances.

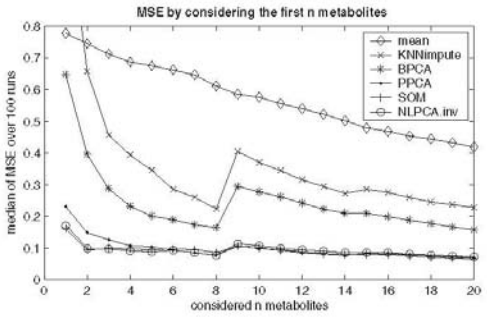


Fig. 6. In contrast to Figure 5 we have considered only the top n metabolites of highest variance, $n = 1, \ldots, 20$, at a fixed missing value rate of 10%. As the dataset contains ratios, a metabolite with a high variance is assumed to be important. The results differ from those in Figure 5. Here, BPCA gives no very good result, but still better than KNNimpute ($k = 10$ neighbours). The best result of PPCA was given with $k = 5$ components. However, at the first five metabolites, this result could still be outperformed by the non-linear techniques, the inverse NLPCA and SOM, which perform almost equally well. All techniques show an abrupt rise at the 9th metabolite (citramalic acid), caused by badly distributed data.

value estimation. The SOM also gives a reasonably good estimation, but the linear approaches BPCA and PPCA, as well as the k-nearest neighbour based approach KNNimpute, fail with this non-linear dataset.

4.2 Missing data: metabolite data

The performance of the missing value estimation techniques was also assessed using a real experimental dataset. For that we used a completely available set of 140 metabolites from our cold stress experiment, see section 5 for more details. Different percentages of values were randomly removed and regarded as missing for the estimation techniques. A good overall missing value estimation is obtained for up to 50 percent missing values. This unexpectedly high rate might be caused by the high redundancy in the data, possibly due to high connectivity or dependency among the metabolites. By comparing the different techniques, we first found that BPCA gives the best average over all 140 metabolites, (Fig. 5). But instead of a good average we are interested in a good estimation of the most important metabolites. As our data values are ratios, see section 5.1, a high variance indicates an important metabolite. Therefore, we compared the performance on the first n metabolites of highest variance which mostly also show a strong non-linear behaviour. Now the results are different, (Fig. 6). The inverse NLPCA and SOM, which perform almost equally well, give the best result at the first five most important metabolites, and perform almost as equally well as the result of PPCA with the remaining metabolites.

4.3 Missing data: gene expression data

To obtain a fair and comprehensive comparison, we also tested the performance of the missing data estimation using a larger set of gene expression data obtained from the same cold stress experiment. The data were again transformed to $\log_2$ ratios, relative to the median of control samples at time zero. In total, 16 996 genes were reduced to 1000 of highest log ratio variance. These genes are expected to be most important as they show the largest relative expression change. Twenty-one samples were measured at seven different time points.

Again, instead of a good averaged missing value estimation over all genes, we are interested in a good estimation of the most important genes, those of highest relative variance. Therefore, the cumulative mean square error (MSE) for the first 30 genes of highest ratio variance is shown (Fig. 7). The results differ from those on the metabolite dataset in Figure 6. All methods give quite similar, but significantly better, results than naive substitution by the mean of the residual values of each gene. However, BPCA which was developed for this kind of high-dimensional datasets, gave the best result for both the averaged estimation (not shown) and the estimation for the first n genes as shown in Figure 7. BPCA is successful because it uses principal components in the lower dimensional data space given by the small number of samples and not by the genes. Similar results can therefore also be obtained by the similar technique of PPCA when applied to the transposed dataset. However, the advantage of BPCA is that no parameter k, the number of used components, has to be chosen as is necessary with PPCA. The results of NLPCA were also improved when applied to the transposed matrix, and with the use of more than one non-linear component ($k = 4$). However, there might be no advantage of a non-linear technique applied to the transposed dataset as a non-linear data structure in gene data space does not necessarily lead to a non-linear structure in sample space (where genes are data points).

Consequently, for estimating missing values in large gene expression datasets BPCA is a good choice. In datasets with a smaller number of variables, as is typical for metabolite or protein datasets, other methods are more suitable. These include non-linear techniques, such as NLPCA or SOM, when the data are non-linearly distributed. Both the gene expression and metabolite datasets, are

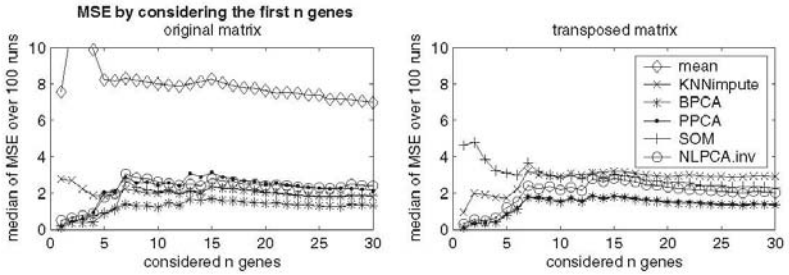


Fig. 7. Missing data algorithms applied to gene expression data of 1000 genes with 10% randomly removed values. The results differ from those on metabolite data in Figure 6. Again, we consider the most important genes of highest ratio variance. The cumulative MSE is given for the first 30 genes of highest ratio variance. All algorithms give significantly better results than the naive substitution by mean. The best result, though, is given by BPCA. Right: the results of most methods can be improved when applied to the transposed matrix. PPCA with $k=5$ components is then almost as good as BPCA, which was applied alone without transposition because it has already an internal transposition.

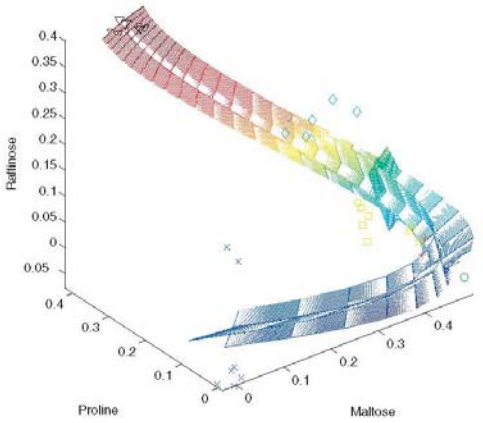


Fig. 8. The first three extracted non-linear components are plotted into the space, given by the top three metabolites of highest variance. The grid represents the new coordinate system after the non-linear transformation. The main curvature, the first non-linear component, shows the trajectory over time in the cold stress experiment. The additional second and third components represent only the noise in the data, but they are useful for regulating the first component.

available at http://nlpca.mpimp-golm.mpg.de. However, our major objective is to detect non-linear components in incomplete datasets. As these components should explain the experimental factors in the data space given by genes (where samples are data points) a transposed matrix is of no use.

5 APPLICATION

The proposed inverse NLPCA model was used to analyse the metabolite response of *A.thaliana* to cold stress at 4°C. This gives us an approximation of the mapping function from a given time point t_i

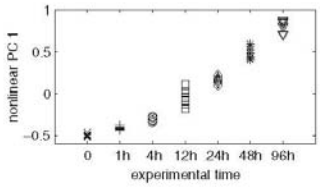


Fig. 9. The extracted first non-linear component represents the time factor. This relation is shown by plotting the first component against the observed experimental time.

to the metabolite responses x, and hence we obtain a 'noise-free' model of the biological cold stress response.

5.1 Data acquisition

We have used gas chromatography/mass spectrometry (GC/MS) to measure 497 metabolites at seven different time points, at 0,1, 4,12,24,48 and 96 h, time point zero represents the control samples; Only 140 metabolites had available measurements for all samples, these metabolites were used in the previous section 4.2 to test the different methods for missing value estimation. In this experimental section the inverse NLPCA is applied to all metabolites which have <1/3 missing values. After removing 109 metabolites, the final dataset contains 388 metabolites (140 complete, 248 incomplete) and 52 samples at seven different time points (7–8 samples per time point).

The data are transformed to log fold changes (log ratios). All measurements of each metabolite $x_i=(x_i^1,\ldots,x_i^{52})^T$ are divided by the median of the control samples at time point zero. Consequently, we are analysing ratios of metabolite concentrations with respect to a control time point. The logarithm $\log_2$ is used to get symmetric changes: $x_{\text{normed}}=\log_2\left(\frac{x}{\text{median}(x_{\text{control}})}\right)$.

5.2 Model parameters

As inverse NLPCA model, we have used a network with a [3-20-388] architecture. This means we have extracted three non-linear components; 20 non-linear hidden units were used to perform the

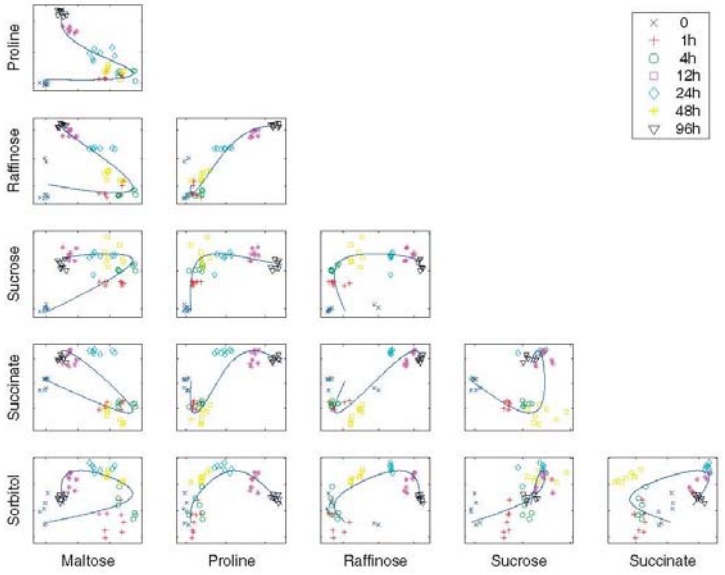


Fig. 10. Scatter plot of six selected metabolites of highest relative variance. The extracted time component (non-linear PC 1) is marked by a curve, which shows a strong non-linear behaviour.

non-linear transformation, and 388 metabolites were approximated. The training was done in 300 iterations. To limit the complexity of the model we also added a weight decay term to the error function $E_{\text{total}} = E + \nu(\sum_i w_i^2 + \sum_j z_j^2)$ with $\nu = 0.001$ and we have extracted the second and third component in a hierarchical order (Scholz and Vigário, 2002), which stabilizes the first component.

The inverse NLPCA model gives us a non-linear transformation from three estimated non-linear components to a 388 dimensional metabolite dataset. This is shown in Figure 8 for the top three metabolites of highest variance.

5.3 Results

The extracted first non-linear component is directly related to the experimental time factor, see Figure 9. This means that the global or main information, represented by variance, is the metabolite change over time. This time trajectory clearly has a non-linear behaviour, see Figure 10. The time component gives a strong curve in the original metabolite data space. It can be seen as a noise-reduced representation of the cold stress response. The inverse model gives us a mapping function $\mathcal{R}^1 \rightarrow \mathcal{R}^{388}$ from a time point t to the response x of all considered 388 metabolites $x = (x_1, \ldots, x_{388})^T$. Thus, we can analyse the approximated response curves for each metabolite, shown in Figure 11. The cold stress is reflected in almost all metabolites; however, the response behaviour is quite different. Some metabolites have a very early positive or negative response, e.g. maltose and raffinose, whereas other metabolites show only a moderate increase.

In classical PCA we can select the metabolites that are most important to a specific component by a rank order of the absolute values from the corresponding eigenvector, also termed loadings or weights. As the components are curves in non-linear PCA, no global ranking is possible. The rank order is different for different positions on the curved component, hence different at different time points in our case. However, we can give a rank order for each individual time point by computing the gradient $q_i = \frac{dx_i}{dt}$ on the non-linear time curve at this time point. The rank order of the top 20 metabolites is shown in Table 2 for an early time point t_1 and a late time point t_2. The influence values $\hat{q}_i$ are the l_2-normalized gradients q_i, $\sum_i (\hat{q}_i)^2 = 1$. The gradient curves over time are shown in Figure 11. We found that even at the last time point of the experiment, 96 hours, there are still some metabolites with significant changes in their concentrations.

6 CONCLUSIONS

NLPCA was achieved by an inverse neural network model that was applicable to incomplete datasets. With this inverse NLPCA we were able to extract non-linear (curved) components from datasets with a large number of missing values. These extracted components can be used, together with the model, to reconstruct the original data, including the missing values. We have shown that in the case of non-linearly structured datasets, both non-linear techniques, the inverse NLPCA and SOM, can improve the missing value estimation performance on the most important metabolites. We have shown that in the case of non-linearly structured datasets, both non-linear techniques, the inverse NLPCA and SOM, can improve the missing value estimation performance for the most important metabolites of the lower dimensional metabolite dataset. In the

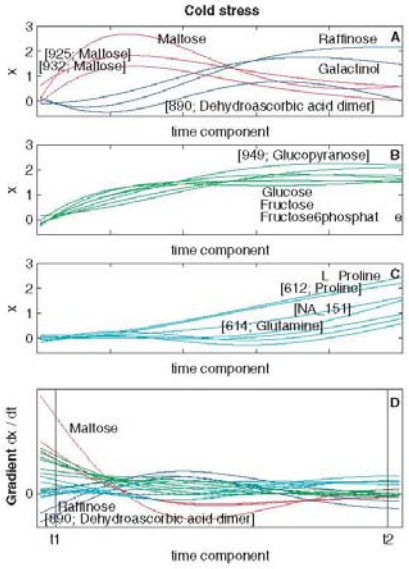


Fig. 11. The top three figures show the different shapes of the approximated metabolite response curves over time. (**A**) Early positive or negative transients, (**B**) increasing metabolite concentrations up to a saturation level, or (**C**) a delayed increase, and still increasing at the last time point. (**D**) The gradients give us the influence of all metabolites at any time point, analogous to loading factors in PCA. A high positive or high negative gradient would be interesting. There is a strong early dynamic, which is quickly moderated, except for some metabolites that are still not stable at the end. Plotted are the top 20 metabolites with the highest absolute gradients. The rank order for the marked early time t_1 and late time t_2 is given in Table 2.

larger gene expression dataset the best missing data estimations were obtained by BPCA and PPCA.

Applied to our cold stress experiment, the first non-linear component was directly related to the experimental time factor. Thus, the inverse NLPCA model gives us the continuous metabolite response over the time frame of the experiment. This trajectory over time is helpful to get a better understanding of the cold stress response. For each time point, including interpolated time points, we are able to give a ranked list of the most important metabolites, analogous to a global ranking in PCA.

The cold stress response clearly showed a non-linear behaviour over time, at the metabolite level (Kaplan *et al.*, 2004). A similar non-linear behaviour was also found in gene expression data from the same cold stress experiment (data not shown). This non-linear analysis can therefore be done in the same way for such data.

Non-linearities are not restricted to temporal experiments, they can also be caused by other continuously changing factors, e.g. different temperatures at a fixed time point. Even natural phenotypes often take the form of a continuous range (Fridman *et al.*, 2004), where non-linearities could exist.

Table 2. Top 20 metabolites at time points t_1 and t_2

t_1, approx. 0.5 h		t_2, approx. 96 h	
$\tilde{q}$	metabolite	$\tilde{q}$	metabolite
0.43	Maltose methoxyamine	0.24	[614; Glutamine]
0.23	[932; Maltose]	−0.20	[890; Dehydroascorbic acid dimer]
0.21	Fructose methoxyamine	0.18	[NA_293]
0.19	[925; Maltose]	0.18	[NA_201]
0.19	Fructose-6-phosphate	0.17	[NA_351]
0.17	Glucose methoxyamine	0.16	[NA_151]
0.17	Glucose-6-phosphate	0.16	L-Arginine
0.16	[674; Glutamine]	0.16	L-Proline
0.16	[NA_1]	−0.14	Sorbitol
0.15	[NA_154]	−0.13	4-Aminobutyric acid
0.14	[NA_341]	0.13	[612; Proline]
0.14	[NA_19]	0.12	[NA_42]
0.14	L-Arginine	−0.11	[NA_118]
0.13	Glycine	−0.11	[NA_37]
0.13	[NA_160]	−0.11	[NA_70]
0.12	[949; Glucopyranose]	0.11	[529; Indole-3-acetic acid]
0.12	[NA_84]	0.10	[NA_210]
−0.12	[890 Dehydroascorbic acid dimer]	0.10	[NA_68]
0.12	[880; Maltose methoxyamine]	−0.10	Galactinol
0.12	L-Glycerol-3-phosphate	−0.10	[NA_117]

The most important metabolites are given for an early time point t_1 of around 0.5 h (interpolated) cold stress and a very late time point t_2 of around 96 h.
The metabolites are ranked by their influences at a specific time point, given by the gradient of the non-linear time component at this time point. As expected maltose, fructose and glucose give a strong early response to cold stress; however, even after 96 hours there are still some metabolites with significant changes in their activity. Brackets '[...]' denote an unknown metabolite, e.g. [925; Maltose] denotes a metabolite with high mass spectral similarity to maltose.

ACKNOWLEDGEMENT

We would like to thank John Lunn for helpful comments on the manuscript. Funding to pay the Open Access publication charges for this article was provided by the Max Planck Society.

Conflict of interest: none declared.

REFERENCES

Bishop,C. (1995) Neural Networks for Pattern Recognition. Oxford University Press, .
Bishop,C. (1999) Variational principal components. *Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99* pp. 509–514.
DeMers,D. and Cottrell,G. W. (1993) Nonlinear dimensionality reduction. In Hanson,D., Cowan,J. and Giles,L. (eds), *Advances in Neural Information Processing Systems 5*. Morgan Kaufmann, San Mateo, CA, pp. 580–587.
Diamantaras,K. and Kung,S. (1996) Principal Component Neural Networks. Wiley, NY.
Fridman,E. *et al.* (2004) Zooming in on a quantitative trait for tomato yield using interspecific introgressions. *Science*, **305**, 1786–1789.
Hassoun,M. H. and Sudjianto,A. (1997) Compression net-free autoencoders. *Workshop on Advances in Autoencoder/Autoassociator-Based Computations at the NIPS 97 Conference*.
Hastie,T. and Stuetzle,W. (1989) Principal curves. *J. American Statistical Association*, **84**, 502–516.
Haykin,S. (1998) Neural Networks-A Comprehensive Foundation, 2nd edn. Prentice Hall.

Hecht-Nielsen,R. (1995) Replicator neural networks for universal optimal source coding. *Science*, **269**, 1860–1863.

Hestenes,M.R. and Stiefel,E. (1952) Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, **49**, 409–436.

Honkela,A. and Valpola,H. (2005) Unsupervised variational bayesian learning of nonlinear models. To appear in *Advances in Neural Information Processing Systems 17*.

Hsieh,W.W. (2004) Nonlinear multivariate and time series analysis by neural network methods. *Reviews of Geophysics*, **42**, RG1003, doi:10.1029/2002RG000112.

Jolliffe,I.T. (1986) Principal Component Analysis. Springer-Verlag, NY.

Kaplan,F. *et al.* (2004) Exploring the temperature-stress metabolome of *arabidopsis*. *Plant Physiology*, **136**, 4159–4168.

Kirby,M. J. and Miranda,R. (1996) Circular nodes in neural networks. *Neural Computation*, **8**, 390–402.

Kohonen,T. (2001) Self-Organizing Maps, 3rd edn, Springer.

Kramer,M. A. (1991) Nonlinear principal component analysis using auto-associative neural networks. *AIChE Journal*, **37**, 233–243.

Lappalainen,H. and Honkela,A. (2000) Bayesian nonlinear independent component analysis by multi-layer perceptrons. In Girolami,M. (ed.), *Advances in Independent Component Analysis*. Springer-Verlag, pp. 93–121.

Little,R.J.A. and Rubin,D.B. (2002) Statistical Analysis with Missing Data. John Wiley & Sons, NY, second edition.

Malthouse,E. C. (1998) Limitations of nonlinear pca as performed with generic neural networks. *IEEE Transactions on Neural Networks*, **9**, 165–173.

Monahan,A.H. *et al.* (2003) The vertical structure of wintertime climate regimes of the northern hemisphere extratropical atmosphere. *J. Climate*, **16**, 2005–2021.

Oba,S. *et al.* (2003) A bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, **19**, 2088–2096.

Oh,J.-H. and Seung,H. (1998) Learning generative models with the up-propagation algorithm. In Jordan,M.I., Kearns,M.J. and Solla,S.A. (eds), *Advances in Neural Information Processing Systems, volume 10*. The MIT Press, pp. 605–611.

Press,W.H, Teukolsky,S.A., Vetterling,W.T. and Flannery,B. P. (1992) Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, second edition.

Raiko,T. and Valpola,H. (2001) Missing values in nonlinear factor analysis. In *Proc. of the 8th Int. Conf. on Neural Information Processing (ICONIP'01)*. Shanghai, pp. 822–827.

Roweis,S. (1997) Algorithms for PCA and SPCA. In *Neural Information Processing Systems 10 (NIPS'97)*. pp. 626–632.

Roweis,S.T. and Saul,L.K. (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**, 2323–2326.

Saul,L.K. and Roweis,S.T. (2004) Think globally, fit locally: unsupervised learning of low dimensional manifolds. *J. Mach. Learn. Res.*, **4**, 119–155.

Schölkopf,B. *et al.* (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, **10**, 1299–1319.

Scholz,M. and Vigário,R. (2002) Nonlinear PCA: a new hierarchical approach. In Verleysen,M. (ed.), *Proceedings ESANN*. pp. 439–444.

Simeka,K. and Kimmel,M. (2003) A note on estimation of dynamics of multiple gene expression based on singular value decomposition. *Math. Biosci.*, **182**, 183–199.

Tenenbaum,J.B. *et al.* (2000) A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, 2319–2323.

Troyanskaya,O. *et al.* (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Verbeek,J.J., Vlassis,N. and Kröse,B. (2002) Procrustes analysis to coordinate mixtures of probabilistic principal component analyzers. Technical report, Computer Science Institute, University of Amsterdam, The Netherlands.

Zhou,X. *et al.* (2003) Missing-value estimation using linear and non-linear regression with bayesian gene selection. *Bioinformatics*, **19**, 2302–2307.

BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 5 2008, pages 732–737
doi:10.1093/bioinformatics/btn023

Data and text mining

TagFinder for the quantitative analysis of gas chromatography—mass spectrometry (GC-MS)-based metabolite profiling experiments

Alexander Luedemann, Katrin Strassburg, Alexander Erban and Joachim Kopka*

Department Prof. L. Willmitzer, Max Planck Institute of Molecular Plant Physiology, Am Muehlenberg 1, D-14476 Potsdam-Golm, Germany

Received on November 30, 2007; revised on January 9, 2008; accepted on January 12, 2008

Advance Access publication January 19, 2008

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Typical GC-MS-based metabolite profiling experiments may comprise hundreds of chromatogram files, which each contain up to 1000 mass spectral tags (MSTs). MSTs are the characteristic patterns of ~25–250 fragment ions and respective isotopomers, which are generated after gas chromatography (GC) by electron impact ionization (EI) of the separated chemical molecules. These fragment ions are subsequently detected by time-of-flight (TOF) mass spectrometry (MS). MSTs of profiling experiments are typically reported as a list of ions, which are characterized by mass, chromatographic retention index (RI) or retention time (RT), and arbitrary abundance. The first two parameters allow the identification, the later the quantification of the represented chemical compounds. Many software tools have been reported for the pre-processing, the so-called curve resolution and deconvolution, of GC-(EI-TOF)-MS files. Pre-processing tools generate numerical data matrices, which contain all aligned MSTs and samples of an experiment. This process, however, is error prone mainly due to (i) the imprecise RI or RT alignment of MSTs and (ii) the high complexity of biological samples. This complexity causes co-elution of compounds and as a consequence non-selective, in other words impure MSTs. The selection and validation of optimal fragment ions for the specific and selective quantification of simultaneously eluting compounds is, therefore, mandatory. Currently validation is performed in most laboratories under human supervision. So far no software tool supports the non-targeted and user-independent quality assessment of the data matrices prior to statistical analysis. TagFinder may fill this gap.

Strategy: TagFinder facilitates the analysis of all fragment ions, which are observed in GC-(EI-TOF)-MS profiling experiments. The non-targeted approach allows the discovery of novel and unexpected compounds. In addition, mass isotopomer resolution is maintained by TagFinder processing. This feature is essential for metabolic flux analyses and highly useful, but not required for metabolite profiling. Whenever possible, TagFinder gives precedence to chemical means of standardization, for example, the use of internal reference compounds for retention time calibration or quantitative standardization. In addition, external standardization is supported for both compound identification and calibration. The workflow of TagFinder comprises, (i) the import of fragment ion data, namely mass, time and arbitrary abundance (intensity), from a chromatography file interchange format or from peak lists provided by other chromatogram pre-processing software, (ii) the annotation of sample information and grouping of samples into classes, (iii) the RI calculation, (iv) the binning of observed fragment ions of equal mass from different chromatograms into RI windows, (v) the combination of these bins, so-called mass tags, into time groups of co-eluting fragment ions, (vi) the test of time groups for intensity correlated mass tags, (vii) the data matrix generation and (viii) the extraction of selective mass tags supported by compound identification. Thus, TagFinder supports both non-targeted fingerprinting analyses and metabolite targeted profiling.

Availability: Exemplary TagFinder workspaces and test data sets are made available upon request to the contact authors. TagFinder is made freely available for academic use from http://www-en.mpimp-golm.mpg.de/03-research/researchGroups/01-dept1/Root_Metabolism/smp/TagFinder/index.html

Contact: Kopka@mpimp-golm.mpg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online and within the TagFinder download from the above URL.

1 INTRODUCTION

Gas chromatography hyphenated to mass spectrometry (GC-MS) is one of the most versatile and widely applied technology platforms in modern metabolomic and fluxomics studies. Metabolic phenotyping has become an integral part of molecular physiology and functional genomics (e.g. Bino *et al.*, 2004; Nicholson *et al.*, 1999; Nielsen and Oliver, 2005; Stephanopoulos *et al.*, 2004; Sumner *et al.*, 2003; Trethewey *et al.*, 1999). As a consequence initiatives for the standardization of high-throughput metabolite analyses have been initiated (Castle *et al.*, 2006; Fiehn *et al.*, 2006; Jenkins *et al.*, 2004; Lindon *et al.*, 2005; Spasić *et al.*, 2006). Guidelines are now available from 10 articles of a recent issue of the Metabolomics journal (Vol. 3, 2007) which comprehensively cover aspects of metabolomics standardization (e.g. Fiehn *et al.*, 2007; Sumner *et al.*, 2007).

*To whom correspondence should be addressed.

However, contrary to the efforts of unifying the standards for data retrieval, mining and interpretation, highly diverse and in part specialized software solutions for the GC-MS data pre-processing have been published. The pre-processing for metabolomic studies typically deals with the seemingly simple task of transforming chemical compound information and respective peak lists from series of chromatography data files into numeric data matrices, which are then amenable to statistical analysis. Both automated peak extraction, and the automated deconvolution of mass spectra allow the comprehensive and non-biased analysis of GC-MS experiments. Previous knowledge about all potential compounds, which may occur in a sample, is not required. In addition, deconvolution reconstructs mass spectra and, thus, provides in principle relevant mass spectra for compound identification. Thus, deconvolution has been an intensely explored topic. In the following, we will shortly summarize basic categories of pre-processing.

The conventional approach of mass spectral deconvolution is based on information present within single chromatograms. Just to mention one early approach, Pool and co-workers developed a backfolding procedure for the mathematical enhancement of GC-MS-based chromatographic curve resolution (Pool *et al.*, 1996, 1997a, b). In contrast, multivariate curve resolution (MCR) and its successor the hierarchical multivariate curve resolution (HDA) may represent the most advanced pre-processing tool. Information from multiple aligned chromatograms is utilized for joined deconvolution (Jonsson *et al.*, 2004, 2005, 2006). The benefit of this multi-chromatogram procedure is the separation of co-eluting MSTs based on the independence of concentration changes of different compounds in many samples. Even mixtures of exactly co-eluting ambient ^{12}C- and fully stable isotope labelled ^{13}C-mass isotopomers can be resolved using MCR (Kopka J, personal communication). However, MCR is highly sensitive to the selection and number of co-analysed chromatogram files and does not allow targeted retrieval of selected fragment ions or extraction of mass isotopomer distributions for flux analysis.

Perhaps, the most widely spread tool may be the mass spectral deconvolution and identification system (AMDIS; http://chemdata.nist.gov/mass-spc/amdis/overview.html; Halket *et al.*, 1999; Stein, 1999), which is provided with the standard mass spectral search and comparison software NIST05 (National Institute of Standards and Technology, Gaithersburg, MD, USA; http://www.nist.gov/srd/mslist.htm). AMDIS was initially designed for purely qualitative analysis, but now also extracts quantitative information, such as the base peak intensity and an estimation of the total intensity of each deconvoluted MST. The main commercial competitor of the universally applicable AMDIS tool is the ChromaTof software (LECO, St. Joseph, MI, USA; http://www.leco.org/), which is exclusive for the GC-EI-TOF-MS and two-dimensional GCxGC-EI-TOF-MS instruments of the vendor. ChromaTof software is designed for in-line acquisition and analysis of GC-TOF-MS chromatograms. The software is customized for the high frequency of mass spectral acquisition (10–500 scans s^{-1}) and the resulting large file sizes, which are generated by fast scanning TOF technology (e.g. Dalluge *et al.*, 2002a, b; Vreuls *et al.*, 1999). Automated mass spectral deconvolution appears to be highly successful for compound discovery, but comes at the price of software errors, such as partially deconvoluted MSTs, mixed or in other words chimeric MSTs, occurrence of artificial MSTs due to electronic noise, and erroneous MST duplications.

For the improvement of data analysis and retrieval from metabolite profiling experiments, software projects were initiated in academia, which were based on the comprehensive extraction of mass selective peak apex intensities. This approach is computationally less demanding compared to the, traditionally preferred, extraction of peak areas. A typical example of these software tools is MetAlign (http://www.pri.wur.nl/UK/products/MetAlign/). MetAlign was initially commercialized and is now free for academic use. It was successfully targeted at supporting LC-MS analyses and includes highly parameterized smoothing, baseline correction and statistical chromatographic alignment options. These options were recently extended by a mass alignment procedure for the accommodation of high mass-accuracy instruments (America *et al.*, 2006; Bino *et al.*, 2005; De Vos *et al.*, 2007; Keurentjes *et al.*, 2006; Vorst *et al.*, 2005). MetAlign is applicable to GC-EI-TOF-MS analysis (Tikunov *et al.*, 2005), but appears not to perform deconvolution, in other words the combination of extracted mass tags into full MSTs. Further software tools allow both non-targeted as well as compound-targeted GC-MS analysis, for example the tools, XCMS (Smith *et al.*, 2006), MathDAMP (Baran *et al.*, 2006), MetaQuant (Bunk *et al.*, 2006), the MSFACTs (Duran *et al.*, 2003; http://www.noble.org/PlantBio/MS/MSFACTs/MSFACTs.html) and the respective refinement, MET-IDEA Broeckling *et al.*, 2006; (http://www.noble.org/Plantbio/MS/MET-IDEA/index.html) or the progressive peak clustering approach (De Souza *et al.*, 2006). First attempts have also been made at compound-targeted processing of GCxGC-EI-TOF-MS files (e.g. Sinha *et al.*, 2004). One unique tool, the BinBase (Fiehn *et al.*, 2005; http://fiehnlab.ucdavis.edu/projects/binbase_setupx/), combines GC-EI-TOF-MS data analysis with a database application, which collects and archives extracted MSTs. Other MST and mass spectral libraries were made publicly available (e.g. Kopka *et al.*, 2005; Schauer *et al.*, 2005; Wagner *et al.*, 2003; http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html) or may be commercially obtained, such as the Wiley library (http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470047860.html) or the NIST05 collection, mentioned above (Ausloos *et al.*, 1999; Halket *et al.*, 2005; http://chemdata.nist.gov/mass-spc/Srch_v1.7/index.html).

In summary, a rich and diverse resource of freely or commercially available tools exists for the pre-processing of GC-MS chromatography data. Each tool has a specific set of parameter settings, which needs to be optimized by human intervention. All tools will in general provide correct and repeatable processing results. But results from different tools are currently difficult to compare. Moreover, most tools do not retrieve mass isotopomer distributions for flux analysis. TagFinder represents the solution to data retrieval for flux studies and allows the comparative analysis of pre-processing procedures. In the following, we will shortly describe the generalized workflow of a TagFinder supported analysis of a GC-(EI-TOF)-MS profiling experiment.

2 WORKFLOW

2.1 Data import

TagFinder software (Supplementary file 1) is a single user application for personal computer systems based on the Java™ programming language and Java™ runtime environment 1.5 or higher (http://www.java.com/de/download/). TagFinder is based on workspaces, which define the mass range and the decimal precision of the RI system. The RI precision is user-defined and should be based on the speed and rate of data acquisition of the GC-MS experiment under investigation. The recommended data import uses the commonly accepted chromatography interchange format NetCDF, which can be exported from almost any vendors' GC-MS acquisition software. TagFinder may utilize non-processed exports. However, the export of baseline corrected and smoothed NetCDF files is recommended. Smoothing and baseline correction are as a rule best performed by vendor and system specific software applications, because different GC-MS technologies may require specific parameter settings and algorithms, which should be optimized by each vendor. An interface for the use of MetAlign for chromatography data pre-processing is available upon request.

TagFinder generates a peak list which corresponds to each NetCDF file. The name of the peak list file will be identical to the name of the processed NetCDF file and is used for subsequent unambiguous annotation of sample information. The peak list files for import into TagFinder software contain mass fragments, specifically the chromatographic peak apices which are linked to mass, RT and peak height information. The mass, retention- and intensity-ranges may be customized prior to data import. TagFinder provides an additional import option for mass spectral deconvolutions and matching results from the ChromaTof software. In detail, the name of the best mass spectral hit, the respective matching factor, the expected RI and the measured RT are imported. If the RI is calculated through an external software, the pre-calculated RI may be imported rather than RT. Two typical structures of TagFinder import files are shown (Fig. 1). These examples demonstrate a typical ChromaTof deconvolution of a single chromatographic peak (Fig. 1A) and the respective result of the same peak after NetCDF file processing and peak apex extraction (Fig. 1B).

Our reference data set for performance testing comprises 32 NetCDF files of approximately 158.100 KB each, generated on a Pegasus III system (LECO, St. Joseph, MI, USA), which are reduced to 57.000–76.000 KB after baseline correction by the ChromaTof software (Version 1, 2002, Pegasus driver 1.61). ChromaTof deconvolution and processing was 20 min per file and generated 600–1.400 KB peak list files in tab delimited text format (cf. Supplementary file 2). TagFinder required ~120 s per baseline corrected NetCDF file for peak list processing and generated 450–850 KB tab delimited text files (cf. Supplementary file 3). These performance data were estimated using a 2.26 GHz, 2.00 GB RAM single Intel Pentium M processor laptop computer with a Microsoft Windows XP Professional operating system (Version 2002, service pack 2). The subsequent performance data were generated using either the ChromaTof processed or the TagFinder processed compendium of reference peak lists. The minimum imported intensity

Fig. 1. *The data import structure of TagFinder.* The elution time window (±0.3 s) of the compound ribitol (5TMS) from a single chromatogram is shown. Pre-processing was performed using either ChromaTof deconvolution (**A**) or the NetCDF file pre-processing implemented in TagFinder (**B**). ChromaTof software performs mass spectral deconvolution of MSTs, mass spectral matching and RI calculation, whereas NetCDF processing only extracts mass tags. The matched name (LIB_ID) is an analyte identifier taken from GMD (http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html). The mass spectral display is truncated and represents non-normalized abundances. The respective values of m/z 73 are underlined. Typical peak lists are shown in supplementary files 2 and 3.

was set to 25 arbitrary abundance units. Import generates a single file from all initial peak lists. Our TagFinder processed reference compendium of 32 chromatograms generated a 1.36 GB TagFinder file within ~167.1 s, whereas the ChromaTof pre-processing reduced the Tagfinder file to 0.1 GB, which is imported within a few seconds. Tagfinder files, which are substantially larger than 2.04 GB, cannot be handled due to limitations of the operating system and JAVA™ run time environment. However, the intensity threshold for data import can be optimized to use the full TagFinder file size. For example, the above 1.36 GB TagFinder file can be reduced to 0.84 GB or 0.58 GB with a 50 or 100 arbitrary abundance unit cut-off, respectively. Thus larger sets of chromatograms can be simultaneously analysed. The advised intensity limit for the processing of up to 250 GC-TOF-MS chromatogram sets using TagFinder is approx. 100–250 arbitrary units.

2.2 Annotation of sample classes

The annotation of sample classes is performed prior to numerical data processing. Sample classifications may be used for supervised data processing methods and, therefore, respective information is established within TagFinder. In order to avoid erroneous annotation, the complete sample information is provided in a sample header section of the final data matrix (cf. Section 2.7.). Sample information may be entered manually or can be imported from a tab delimited sample annotation file. The required table must have a column labelled 'RAWNAME', which contains the names of all imported data files (cf. Section 2.1.; naming of the peak list files). All other sample classifications and information can be freely defined. A column labelled 'AMOUNT' is suggested, which should contain information on the sample amount or volume

required for subsequent data normalization after TagFinder processing.

2.3 RI calculation

RI calculation is performed using the retention time of internal reference substances, such as *n*-alkanes or fatty acid methyl esters. RI definitions are made by the user. TagFinder searches for the retention times of added internal reference substances using specific, unique, user-defined fragment ions. Single fragment ions or full and partial mass spectra can be used for the queries, which are restricted to customized retention time windows. Ambiguous results are solved by adjustment of the retention time window, user intervention or automated selection of the highest intensity fragment ion within the selected RT window. A result file is generated and stored, which contains the RI definitions, and the corresponding retention times of each NetCDF file. We recommend storage of this retention time file and of the sample annotation file for documentation within the respective TagFinder workspace. The retention time file is used for automated RI calculation using linear interpolation between the retention time anchors (van den Dool and Kratz, 1963). The user supported generation of the retention time file from our TagFinder pre-processed example requires 10–20 min. The RI calculation of this exemplary data set is completed within ~114.1 s. TagFinder pre-processing requires search for single or pairs of unique mass fragments, whereas ChromaTof pre-processing provides full mass spectra, which can be queried by generic mass spectra of the compound class used to calibrate chromatographic retention.

2.4 Mass tag generation

Mass fragments of all chromatography files from an experiment are sorted by mass and calculated RI. Within this chromatography sorted array RI gaps are scanned, which separate mass fragments of equal mass. This process aligns and bins mass fragments of equal mass across all files of an experiment allowing for the technical variability of RI determination. The bins are in the following called mass tags. Mass tags receive the properties minimum, maximum RIs, RI width and median RI and an average intensity. The main selectable parameters for the gap finding process are, (i) the scanning distance between mass tags and (ii) the minimum width of mass tags. In addition the minimum abundance, maximum width of mass tags, number of mass occurrences among all files or within groups of replicate samples can be set.

Exclusions-masses, as well as mass- and RI-windows allow test runs for selected chromatographic regions and mass ranges. Trials prior to the final generation of a data matrix should be performed. The restriction to a narrow RI window is recommended for parameter optimization. Each GC-MS experiment differs with respect to the absolute concentrations and concentration range of hundreds of compounds. Therefore, previously optimized parameter settings should be critically revised and adapted for each new GC-TOF-MS profiling experiment.

In the course of mass tag generation TagFinder allows the selection of either, the maximum, the average or the sum of intensities for aggregation of mass fragments with equal mass within the same peak list file. This aggregation procedure is necessary, because a single peak list and the original chromatogram may contain more than one mass fragment of equal mass within the generated RI window of the mass tag. This observation may be unexpected, but these options had to be implemented to solve the typical deconvolution errors of AMDIS or ChromaTof software. In addition, multiple peak apices may occur, if chromatogram files with a high acquisition rate or technical noise can only be insufficiently smoothed.

2.5 Time group combination

Mass tags are grouped by TagFinder into so-called time groups using the overlap of RI windows. Mass spectra of each time group are then reconstructed for mass spectral matching using—in a first simplified approach—the average intensities of mass tags from all chromatograms of an experiment. All mass tags exhibiting similar RI median are grouped. For this purpose the mass tags are first arrayed according to ascending RI median. Then steep, stepwise increases of the RI median are detected in order to separate consecutive time groups (Fig. 2). The resolving criteria for time groups are the lack of overlap between the median RI of the preceding mass tag compared to the minimum RI of the following.

2.6 Time group clustering

Time groups of complex metabolite profiles typically contain mass tags of multiple co-eluting compounds. In addition, mass tags which are non-specific may occur. Non-specific mass tags result from compounds with similar chemical moieties, which may cause fragments of identical mass. Pearson or Spearman correlation is applied to the intensity vectors of mass tags of the same time group. This procedure finds correlated clusters of mass tags. Significance and the coefficient of correlation can be set to vary the stringency of time group clustering. The clustering approach uses the observation of most mass spectrometric devices, namely that the intensity ratios of mass

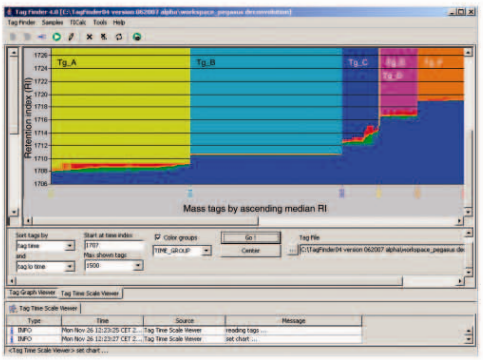


Fig. 2. *Time grouping and clustering.* The screen shot shows the alignment of mass tags by ascending median RI. Green and red bars indicate minimum and maximum RI of each mass tag, the red to green boundary represents median RI. Coloured underlay demonstrates the grouping into time groups (Tg) A F.

fragments representing a single compound are constant and concentration independent within the range of quantification.

Like the reconstruction of mass spectra from time groups the constituent clusters are transformed into mass spectra using average intensities. The size of clusters can be restricted to a minimum number of mass tags per cluster. Mass tags, which are not correlated according to the significance and correlation coefficient thresholds, are maintained within the data matrix for non-biased fingerprinting analysis, but may also be discarded from further analysis.

2.7 Matrix generation

The data matrix generation is performed after time grouping and time group clustering. The matrix contains all initial intensity data, namely the non-normalized abundance data of each observed mass. Sample information is attached to the header section (cf. Section 2.2). Time group and cluster assignments, mass information, RI data and mass spectral matching results, as far as imported from external processing, are attached to the mass tag row information (cf. Supplementary File 4). The matrix is generated as a tab-delimited text file, which can be imported into statistical tools for data normalization, transformation and statistical analysis. We recommend the TM4 multi-experiment viewer (Saeed *et al.*, 2003; Saeed *et al.*, 2006) for a first visual assessment.

2.8 Retrieval of selective mass tags

We use libraries of mass spectra and RI of authenticated reference compounds (Kopka *et al.*, 2005; Schauer *et al.*, 2005; Wagner *et al.*, 2003) for compound identification through the matching of reconstructed mass spectra representing time groups or respective clusters within the exported data matrix (Fig. 3).

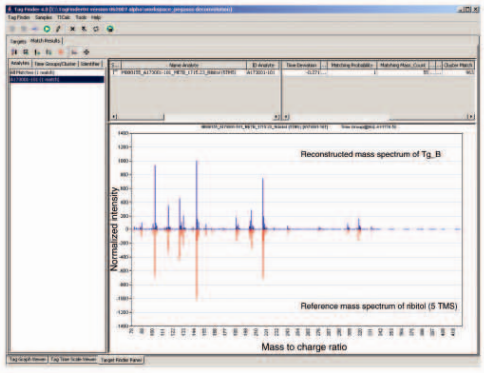


Fig. 3. *The target finding window.* The inset figure shows the head to tail matching of a reference mass spectrum obtained from an authenticated ribitol preparation, bottom (red) and the reconstructed mass spectrum of the respective time group, top (blue). Deviation of retention index (%), number of matching mass tags and the match value are shown.

In addition, the matching can be performed either by using the mass spectral reconstruction from all chromatograms of an experiment or by choice of selected chromatograms, which should contain a known set of targeted reference compounds. We advise to use such reference mixtures for external standardization of each profiling experiment. These reference mixtures, if integrated into the final data matrix, solve ambiguities of compound identification, especially of those chemical isomers, which can only be distinguished by RI and not by mass spectral criteria.

TagFinder extracts all mass tags of time groups based on compound identification by mass spectral matching within expected RI windows. For an improved selectivity, the identified constituent cluster may be obtained and the retrieval of predefined mass fragments is enabled. Single or small sets of selective predefined masses may be useful for comparison of results between data matrices of multiple profiling experiments. The retrieval of targeted fragment masses is also highly useful for mass isotopomer ratio profiling using GC-TOF-MS (Birkemeyer *et al.*, 2005) or for flux analysis as demonstrated by Huege *et al.* (2007), which necessitates the retrieval of mass isotopomer distributions.

3 CONCLUSION

In conclusion, we offer a software tool for the alignment of large GC-MS-based metabolite profiling experiments into statistically accessible data matrices. The matrix generation is directed by co-analysis of RI marker substances within each chromatogram and the simultaneous in-parallel analysis of mixtures of reference compounds is recommended. In addition, we offer automated extraction of quantitative data from predefined mass fragments, time groups of mass fragments or clusters of intensity-correlated mass fragments. This extraction of quantitative data is supported by mass spectral matching to reference mass spectra within preset RI windows as are provided by reference libraries (e.g. Kopka *et al.*, 2005; Schauer *et al.*, 2005). The data matrix generation and matching procedures allow both automation and user intervention for parameter optimization. Thus we present, what we think is the ideal tool for modern metabolomics and fluxomics studies. TagFinder supports non-biased metabolomic fingerprinting, footprinting and profiling experiments (e.g. Kopka *et al.*, 2004, 2006a, b) and, moreover, the metabolite targeted analysis of changes in both metabolite pools and flux.

ACKNOWLEDGEMENTS

The authors acknowledge the long standing support and encouragement by Prof. L. Willmitzer, Max Planck Institute of Molecular Plant Physiology (MPI-MP), Am Muehlenberg 1, D-14476 Potsdam-Golm, Germany. We thank Prof. J. Selbig, University of Potsdam, D-14476 Potsdam-Golm, Germany and Dr D. Walther, MPI-MP, for fruitful discussions. This work was supported by the Max Planck Society, the Bundesministerium für Bildung und Forschung (BMBF), grant PTJ-BIO/0312854 and the European META-PHOR project, FOOD-CT-2006-036220.

Conflict of Interest: none declared.

REFERENCES

America,A.H.P. *et al.* (2006) Alignment and statistical difference analysis of complex peptide data sets generated by multidimensional liquid chromatography mass spectrometry. *Proteomics*, **6**, 641 653.

Ausloos,P. *et al.* (1999) The critical evaluation of a comprehensive mass spectral library. *J. Am. Soc. Mass Spectrom.*, **10**, 287 299.

Baran,R. *et al.* (2006) MathDAMP: a package for differential analysis of metabolite profiles. *BMC Bioinformatics*, **7**, 530.

Bino,R.J. *et al.* (2004) Potential of metabolomics as a functional genomics tool. *Trends Plant Sci.*, **9**, 418 425.

Bino,R.J. *et al.* (2005) The light-hyperresponsive high pigment-2 mutation of tomato: alterations in the fruit metabolome. *New Phytol.*, **166**, 427 438.

Birkemeyer,C. *et al.* (2005) Metabolome analysis: the potential of *in vivo* labeling with stable isotopes for metabolite profiling. *Trends Biotechnol.*, **23**, 28 33.

Broeckling,C.D. *et al.* (2006) MET-IDEA: Data extraction tool for mass spectrometry-based metabolomics. *Anal. Chem.*, **78**, 4334 4341.

Bunk,B. *et al.* (2006) MetaQuant: a tool for the automatic quantification of GC/MS-based metabolome data. *Bioinformatics*, **22**, 2962 2965.

Castle,A.L. *et al.* (2006) Metabolomics Standards Workshop and the development of international standards for reporting metabolomics experimental results. *Briefings Bioinformatics*, **7**, 159 165.

Dalluge,J. *et al.* (2002a) Optimization and characterization of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometric detection (GC × GC-TOF MS). *J. Sep. Sci.*, **25**, 201 214.

Dalluge,J. *et al.* (2002b) Comprehensive two-dimensional gas chromatograpy with time-of-flight mass spectrometric detection applied to the determination of pesticides in food extracts. *J. Chromatogr. A*, **965**, 207 217.

De Souza,D.P. *et al.* (2006) Progressive peak clustering in GC-MS metabolomic experiments applied to Leishmania parasites. *Bioinformatics*, **22**, 1391 1396.

De Vos,R.C.H. *et al.* (2007) Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nature Protocols*, **2**, 778 791.

Duran,A.L. *et al.* (2003) Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics*, **19**, 2283 2293.

Fiehn,O. *et al.* (2005) Automatic annotation of metabolomic mass spectra by integrating experimental metadata. *Proc. Lect. Notes Bioinformatics*, **3615**, 224 239.

Fiehn,O. *et al.* (2006) Establishing reporting standards for metabolomic and metabonomic studies: a call for participation. *Omics- J. Intergrat. Biol.*, **10**, 158 163.

Fiehn,O. *et al.* (2007) The metabolomics standards initiative (MSI). *Metabolomics*, **3**, 175 178.

Halket,J.M. *et al.* (1999) Deconvolution gas chromatography mass spectrometry of urinary organic acids potential for pattern recognition and automated identification of metabolic disorders. *Rapid Commun. Mass Spectrom.*, **13**, 279 284.

Halket,J.M. *et al.* (2005) Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS. *J. Exp. Bot.*, **56**, 219 243.

Huege,J. *et al.* (2007) GC-EI-TOF-MS analysis of in vivo carbon-partitioning into soluble metabolite pools of higher plants by monitoring isotope dilution after ($^{13}CO_2$)-labelling. *Phytochemistry*, **68**, 2258 2272.

Jenkins,H. *et al.* (2004) A proposed framework for the description of plant metabolomics experiments and their results. *Nat. Biotechnol.*, **22**, 1601 1606.

Jonsson,P. *et al.* (2004) A strategy for identifying differences in large series of metabolomic samples analysed by GC/MS. *Anal. Chem.*, **76**, 1738 1745.

Jonsson,P. *et al.* (2005) High-throughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses. *Anal. Chem.*, **77**, 5635 5642.

Jonsson,P. *et al.* (2006) Predictive metabolite profiling applying hierarchical multivariate curve resolution to GC-MS data a potential tool for multi-parametric diagnosis. *J. Proteome Res.*, **5**, 1407 1414.

Keurentjes,J.J.B. *et al.* (2006) The genetics of plant metabolism. *Nat. Genetics*, **38**, 842 849.

Kopka,J. *et al.* (2004) Metabolite profiling in plant biology: platforms and destinations. *Genome Biol.*, **5**, 109 117.

Kopka,J. *et al.* (2005) GMD@CSBDB: The Golm metabolome database. *Bioinformatics*, **21**, 1635 1638.

Kopka,J. (2006a) Current challenges and developments in GC-MS based metabolite profiling technology. *J. Biotechnol.*, **124**, 312 322.

Kopka,J. (2006b) Gas chromatography mass spectrometry. In: Nagata,T., Lörz,H., Widholm,J.M. (eds.) *Biotechnology in agriculture and forestry* Vol. 57: Saito,K, Dixon,RA, Willmitzer,L (eds) Plant metabolomics. Springer-Verlag, Berlin Heidelberg New York, pp. 3 20.

Lindon,J.C. *et al.* (2005) The Consortium for Metabonomic Toxicology (COMET): aims, activities and achievements. *Pharmacogenomics*, **6**, 691 699.

Nicholson,J.K. *et al.* (1999) 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, **29**, 1181 1189.

Nielsen,J. and Oliver,S. (2005) The next wave in metabolome analysis. *Trends Biotechnol.*, **23**, 544 546.

Pool,W.G. *et al.* (1996) Backfolding applied to differential gas chromatography/mass spectrometry as a mathematical enhancement of chromatographic Resolution. *J. Mass Spectrom.*, **31**, 509 516.

Pool,W.G. *et al.* (1997a) Automated extraction of pure mass spectra from gas chromatographic/mass spectrometric data. *J. Mass Spectrom.*, **32**, 438 443.

Pool,W.G. *et al.* (1997b) Automated processing of GC/MS data: quantification of the signals of individual components. *J. Mass Spectrom.*, **32**, 1253 1257.

Saeed,A.I. *et al.* (2003) TM4: A free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374 378.

Saeed,A.I. *et al.* (2006) TM4 microarray software suite. *Methods Enzymol.*, **411**, 134 193.

Schauer,N. *et al.* (2005) GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett.*, **579**, 1332 1337.

Sinha,A.E. *et al.* (2004) Algorithm for locating analytes of interest based on mass spectral similarity in GC x GC-TOF-MS data: analysis of metabolites in human infant urine. *J. Chromatogr. A*, **1058**, 209 215.

Smith,C.A. *et al.* (2006) XCMS: Processing mass spectrometry data for metabolite profiling using Nonlinear peak alignment, matching, and identification. *Anal. Chem.*, **78**, 779 787.

Spasić,I. *et al.* (2006) MeMo: a hybrid SQL/XML approach to metabolomic data management for functional genomics. *BMC Bioinformatics*, **7**, 281.

Stein,S.E. (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *J. Am. Soc. Mass Spectrom.*, **10**, 770 781.

Stephanopoulos,G. *et al.* (2004) Exploiting biological complexity for strain improvement through systems biology. *Nat. Biotechnol.*, **22**, 1261 1267.

Sumner,L.W. *et al.* (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry*, **62**, 817 836.

Sumner,L.W. *et al.* (2007) Proposed minimum reporting standards for chemical analysis. *Metabolomics*, **3**, 211 221.

Tikunov,Y. *et al.* (2005) A novel approach for non-targeted data analysis for metabolomics: large-scale profiling of tomato fruit volatiles. *Plant. Physiol.*, **139**, 1125 1137.

Trethewey,R.N. *et al.* (1999) Metabolic profiling: a Rosetta stone for genomics? *Curr. Opin. Plant Biol.*, **2**, 83 85.

Van den Dool,H. and Kratz,P.D. (1963) A generalization of the retention index system including linear temperature programmed gas-liquid partition chromatography. *J. Chromatogr.*, **11**, 463 471.

Vorst,O. *et al.* (2005) A non-directed approach to the differential analysis of multiple LCMS derived metabolic profiles. *Metabolomics*, **1**, 169 180.

Vreuls,R.J.J. *et al.* (1999) Gas chromatography-time-of-flight mass spectrometry for sensitive determination of organic microcontaminants. *J. Mircocolumn. Sep.*, **11**, 663 675.

Wagner,C. *et al.* (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry*, **62**, 887 900.

Journal of Chromatography B, 871 (2008) 182–190

Contents lists available at ScienceDirect

Journal of Chromatography B

journal homepage: www.elsevier.com/locate/chromb

Retention index thresholds for compound matching in GC–MS metabolite profiling☆

Nadine Strehmel, Jan Hummel, Alexander Erban, Katrin Strassburg, Joachim Kopka *

Max Planck Institute of Molecular Plant Physiology, Department Prof. L. Willmitzer, Am Muehlenberg 1, D-14476 Potsdam-Golm, Germany

ARTICLE INFO

Article history:
Received 7 February 2008
Accepted 14 April 2008
Available online 8 May 2008

Keywords:
Retention index matching
Gas chromatography
GC–MS
Metabolite profiling
Metabolomics

ABSTRACT

The generation of retention index (RI) libraries is an expensive and time-consuming effort. Procedures for the transfer of RI properties between chromatography variants are, therefore, highly relevant for a shared use. The precision of RI determination and accuracy of RI transfer between 8 method variants employing 5%-phenyl-95%-dimethylpolysiloxane capillary columns was investigated using a series of 9 *n*-alkanes (C_{10}–C_{36}). The precision of the RI determination of 13 exemplary fatty acid methyl esters (C_8 ME–C_{30} ME) was 0.22–0.33 standard deviation (S.D.) expressed in RI units in low complexity samples. In the presence of complex biological matrices this precision may deteriorate to 0.75–1.11. Application of the previously proposed Kováts, van den Dool or 3rd–5th order polynomial regression algorithms resulted in similar precision of RI calculation. For transfer of empirical van den Dool-RI properties between the chromatography variants 3rd order regression was found to represent the minimal necessary assumption. The range of typical regression coefficients was $r^2 = 0.9988$–0.9998 and accuracy of RI prediction between chromatography variants varied between 5.1 and 19.8 (0.29–0.69%) S.D. of residual RI error, $RI_{predicted} - RI_{determined}$ ($n > 64$). Accuracy of prediction was enhanced when subsets of chemically similar compound classes were used for regression, for example organic acids and sugars exhibited 0.78 ($n = 29$) and 3.74 ($n = 37$) S.D. of residual RI error, respectively. In conclusion, we suggest use of percent RI error rather than absolute RI units for the definition of matching thresholds. Thresholds of 0.5–1.0% may apply to most transfers between chromatography variants. These thresholds will not solve all matching ambiguities in complex samples. Therefore, we recommend co-analysis of reference substances with each GC–MS profiling experiment. Composition of these defined reference mixtures may best approximate or mimic the quantitative and qualitative composition of the biological matrix under investigation.

1. Introduction

Gas chromatography hyphenated to mass spectrometry (GC–MS) is one of the most versatile and widely applied technology platforms in modern metabolomic and fluxomic studies. Post-genomic molecular physiology increasingly utilises metabolic phenotyping approaches on the quest towards systems biology [1–6]. In recent years standardisation of qualitative and quantitative aspects of these high-throughput analyses has been discussed and minimum laboratory and reporting standards were proposed [7–12]. This study aims to contribute to this ongoing process. We explored the use of retention index (RI) properties for the matching of compound identities in routine GC–MS based metabolite profiling experiments (e.g. [13,14]).

The potential of mass spectral matching to commercial libraries, such as the NIST (http://chemdata.nist.gov/mass-spc/Srch_v1.7/index.html) [15–17] and the Wiley (http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470047860.html) collections was recognised in early profiling studies. Mass spectral matching was shown to be a highly useful and necessary criterion for metabolite identification. However, mass spectral matching alone was found insufficient for non-ambiguous identification, the major obstacle being the presence of multiple structural isomers in highly complex biological samples. As a consequence, RI information based on *n*-alkanes was suggested as an additional supporting criterion for compound matching and recognition (e.g. [18,19]). Moreover, the last update of the NIST05 mass spectral library comprised empirically determined RI information and an implementation of automated RI prediction [20,21]. Subsequently, mass spectral and retention index libraries, which were

☆ This paper is part of a special volume entitled "Hyphenated Techniques for Global Metabolite Profiling", guest edited by Georgios Theodoridis and Ian D. Wilson.

* Corresponding author. Tel.: +49 331 567 8262; fax: +49 331 567 898262.
E-mail address: kopka@mpimp-golm.mpg.de (J. Kopka).

1570-0232/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2008.04.042

dedicated to the analysis of the typically derivatised, methoxymated and trimethylsilylated components of routine metabolite profiling experiments, have been collected [22]. The results of these efforts were made available to the metabolite profiling community through the Golm Metabolome Database (GMD, http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html) [23].

In parallel, a software tool for the pre-processing of GC–MS based profiling experiments was developed [24]. This tool utilises both mass spectral and RI properties for compound identification. With this software tool in place and a dedicated metabolite profiling library available, which comprises contributions from multiple chromatography variants, we investigated both the precision of empirical RI determination and the potential of transferring RI properties between system variants. A previous study on the prediction of RI properties grouped information from either polar or non-polar chromatography systems. Median RI prediction errors of 65 (3.9%) and 46 (3.2%), respectively, were achieved, when chemical group contributions of compounds were considered [21]. On the other hand an initial test within GMD indicated that median predictability may be as good as ±4.81 RI units, if RI was determined using equivalent polar phases [19]. Silylation was reported to mask the functionality of substituent moieties and may allow polyfunctional compounds to revert to a "virtual hydrocarbon state" [25]. Indeed, Kováts indices of silylated compounds were predicted with a typical accuracy below 3% using in a first approximation linear regression functions which considered the atom number of the analytes [25].

Through our study we hope to contribute to the efficient sharing of RI reference libraries, such as GMD, between laboratories and present prerequisite criteria, such as empirical estimations of thresholds for retention index matching. Thus, we hope to support the ultimate goal initiated by the National Institute of Standards and Technology (NIST) to develop and utilise a general data base of chromatographic retention properties for the integrative RI and mass spectral matching of organic compounds [20]. Moreover, the chemometric efforts of predicting RI properties based on feature extractions from molecular structures currently appear to lack high accuracy. Respective predictions and training data sets may now be evaluated considering the information on the maximum possible precision of RI determination and transfer between chromatography variants which are provided through our study.

2. Experimental

2.1. Chemicals for retention time standardisation

Two compound classes are currently in use for the standardisation of retention times in routine GC–MS metabolite profiling experiments, namely *n*-alkanes (e.g. [18,26]) and *n*-alkyl fatty acid methyl esters (e.g. [27]). In this study the following reference substances were obtained from Sigma–Aldrich (Sigma–Aldrich Chemie GmbH, Munich, Germany) at highest available purity; *n*-alkanes: decane (CAS: 124-18-5), dodecane (CAS: 112-40-3), pentadecane (CAS: 629-62-9), octadecane (CAS: 593-45-3), nonadecane (CAS: 629-92-5), docosane (CAS: 629-97-0), octacosane (CAS: 630-02-4), dotriacontane (CAS: 544-85-4) and hexatriacontane (CAS: 630-06-8); *n*-alkyl fatty acid methyl esters: methyl octanoate (CAS: 111-11-5), methyl nonanoate (CAS: 1731-84-6), methyl decanoate (CAS: 110-42-9), methyl dodecanoate (CAS: 111-82-0), methyl myristate (CAS: 124-10-7), methyl palmitate (CAS: 112-39-0), methyl stearate (CAS: 112-61-8), methyl arachidate (CAS: 1120-28-1), methyl behenate (CAS: 929-77-1), methyl tetracosanoate (CAS: 2442-49-1), methyl hexacosanoate (CAS: 5802-82-4), methyl octacosanoate (CAS: 55682-92-3) and methyl melissate (CAS: 629-83-4). Alkanes were dissolved in pyridine at 0.22 mg/mL final concentration [26]. The fatty acid methyl ester (FAME) mixture was prepared separately using chloroform with final concentrations adjusted to 0.4 mg/mL or 0.8 mg/mL for liquid or solid C_8 ME–C_{30} ME standards [27]. Variations of these basic sets of retention marker mixtures are reported below (cf. Section 2.4.10).

Table 1
Defined mixture of 28 authenticated commercially available reference substances (DRM-mix)

Name	Reference number of the chemical abstracts service (CAS)	Final concentration (mg/L)
Benzoic acid, 4-hydroxy-	CAS: 99-96-7	9.9
Alanine, D-	CAS: 338-69-2	10.3
Caffeic acid	CAS: 331-39-5	10.3
Cholesterol	CAS: 57-88-5	10.0
Citramalic acid, D-	CAS: 626-10-8	5.0
Citric acid	CAS: 77-92-9	10.0
Fucose, L-	CAS: 2438-80-4	10.3
Glucose, alpha-D-	CAS: 492-62-6	10.0
Glutaric acid, 2-oxo-	CAS: 328-50-7	10.1
Glycine	CAS: 56-40-6	10.1
Isoleucine, DL-	CAS: 443-79-8	9.9
Lactitol	CAS: 81025-04-9	5.1
Lactose, beta-D-	CAS: 5965-66-2	5.2
Lanosterol	CAS: 79-63-0	10.5
Maltose, D-	CAS: 6363-53-7	20.8
Maltotriose	CAS: 1109-28-0	10.2
Palatinose	CAS: 13718-94-0	10.1
Panthothenic acid, D-	CAS: 137-08-6	9.9
Putrescine	CAS: 333-93-7	9.6
Pyridine, 3-hydroxy-	CAS: 109-00-2	9.9
Ribitol	CAS: 488-81-3	10.0
Ribose, D-	CAS: 50-69-1	10.3
Sorbitol, D-	CAS: 50-70-4	5.3
Sorbose, L-	CAS: 414-273-3850	9.9
Stigmasterol	CAS: 83-48-7	9.6
Threitol, DL-	CAS: 6968-16-7	10.6
Urea	CAS: 57-13-6	10.3
Valine, L-	CAS: 72-18-4	10.1

2.2. Preparations to assess the effect of matrix on retention time standardisation

2.2.1. Defined reference mixture of authenticated substances (DRM-mix)

A defined mixture of 28 authenticated reference substances referred to as the defined reference material (DRM-mix) was prepared as suggested by Fiehn and co-workers on the 2nd annual conference of the Metabolomics Society, 2006 in Boston, MA, USA (http://fiehnlab.ucdavis.edu/Boston%202006%20workshop.pdf). The substances were dissolved in chloroform–methanol–water (1:2.5:1, v/v/v) and diluted to 1 L final volume (Table 1). A defined volume of DRM, 267 μL, was dried using a VR Maxi vacuum concentrator with rotors R96-13 or R120-111 (Jouan Nordic, Allerod, Denmark) fitted to a hold-back vacuum pump (Ilmvac GmbH, Ilmenau, Germany) and subsequently subjected to routine chemical derivatisation and GC–MS profiling analysis (cf. Sections 2.3 and 2.4).

2.2.2. Defined reference material of a yeast intracellular extract (DRM-yeast)

A 1 L liquid batch of yeast, *Saccharomyes cerevisiae* strain S288C, was cultivated from a deep frozen stock using synthetically defined growth medium supplemented with yeast nitrogen base (Difco, Kansas City, MO, USA). Cells were harvested at optical density (OD_{595}) ~1.8. Subsequently intracellular metabolites were prepared as described earlier [26]. In short, 5 mL of yeast culture was

rapidly mixed with 20 mL methanol–water (6:4, v/v), which was pre-cooled to 60 °C. The metabolically inactivated cells were separated from residual growth medium and surplus methanol–water by centrifugation at <−20 °C. Cell pellets were re-suspended and extracted (1) 15 min at 70 °C using 374 μL extraction medium comprising 350 μL methanol, 12 μL internal standard 1 and 12 μL internal standard 2 (cf. below), (2) 10 min at 30 °C using 263 μL chloroform–water (188:75, v/v). The two respective supernatants after centrifugation were combined without liquid partitioning. Extracts of multiple preparations were pooled and equal 500 μL aliquots were dried by vacuum concentration for subsequent GC–MS profiling analysis (cf. Sections 2.3 and 2.4). This defined reference material is named DRM-yeast. The internal standards 1 and 2 are part of the routine metabolite preparation procedure [26], but were not required for this study. For the purpose of complete reporting, internal standard 1 contained 0.2 mg/mL ribitol (CAS: 488-81-3), 1 mg/mL 2,3,3,3-d_4-alanine (CAS: 53795-92-9), and 0.5 mg/mL D-isoascorbic acid (CAS: 89-65-6) dissolved in methanol and water, respectively. Internal standard 2 consisted of methyl nonadecanoate (CAS: 1731-94-8) dissolved at 2 mg/mL in chloroform.

2.2.3. Defined reference material of a rice leaf extract (DRM-rice)

Rice seeds, *Oryza sativa* ssp. indica, were submersed for 60 s in warm water (40 °C), transferred to Petri dishes containing wet cellulose tissue and germinated in the dark. After 2 days germinating seedlings were acclimated to the illuminated greenhouse and growth was allowed to continue to 3–5 cm seedling size. Subsequently, rice seedlings were transferred to hydroponic culture with a weekly exchange of liquid medium [28]. Four weeks after transfer complete shoot material was harvested and shock-frozen in liquid nitrogen. A pooled sample of shoot material from 25 plants was ground under liquid nitrogen to obtain a fine homogenous powder. An aliquot of 120 mg from this homogenate was extracted 15 min at 70 °C with 300 μL methanol, and 30 μL of internal standard 1 and internal standard 2, respectively. Finally 600 μL chloroform–water (1:2, v/v) was added, liquid phase partitioning performed by centrifugation. Multiple preparations of the upper polar phase were pooled and equal 40 μL aliquots dried by vacuum concentration. This defined reference material is in the following called DRM-rice.

2.3. Synthesis of analytes by chemical derivatisation

The dried materials were re-dissolved and chemically modified by 90 min agitation at 30 °C with 10 μL methoxyamine reagent, i.e. 40 mg/mL methoxyamine hydrochloride (CAS: 593-56-6) in pyridine. Then 90 μL reagent mixture, comprising *N*-methyl-*N*-trifluoroacetamide (MSTFA, CAS: 24589-78-4) trimethylsilylation reagent, *n*-alkane-mixture, and FAME-mixture (1000:16:4, v/v/v) were added and agitation continued 30 min at 37 °C. Injection volume was 1 μL of 100 μL final derivative volume [27]. In the following we define the term, analyte, to represent the products of chemical derivatisation, which are subjected to GC–MS analysis. A single compound may generate multiple analytes due to partial silylation and/or E,Z-isomers formed by methoxymation.

2.4. Chromatography variants

In this study we analysed a library collection [22,23] of analyte RI properties, which were recorded in various laboratories using essentially 8 variants of the original GC–MS based metabolite profiling method [13,14]. Besides the use of three detector technologies, namely quadrupole (Q), ion trap (TRAP) and time of flight (TOF) based mass spectral detection, which were deemed irrelevant for the present investigation, chromatography settings were modified. Specifically temperature programming, type of capillary column and choice of column manufacturer were varied. Most chromatography variants used 5%-phenyl-95%-dimethylpolysiloxane (5PDM) or equivalent capillary columns. For comparative purposes we included a variant using 35%-phenyl-65%-dimethylpolysiloxane (35PDM). In the following we describe the essential parameter selections of our own chromatography variants and present the relevant settings as extracted from publications of the other contributing laboratories.

2.4.1. Variant 1 (5PDM_VF5_9_TOF)

Variant 1 [26] uses helium carrier gas at 1 mL/min under constant flow control. Splitless injection at 230 °C was performed with flow transiently reduced to 0.6 mL/min into a conical, single taper liner with deactivated glass wool (Agilent Technologies, Böblingen, Germany). Purge time and flow reduction was 1 min. The 6890N gas chromatography system (Agilent Technologies) was mounted with a 5PDM VF-5 ms, 0.25 μm film thickness, 30 m × 0.25 mm fused silica capillary column (Varian, Darmstadt, Germany), which had an integrated 10 m guard column. The temperature programming comprised an initial 1 min isothermal period at 70 °C, a 9 °C/min ramp to 350 °C and a final 5 min constant heating at 350 °C. TOF-detection was performed using a Pegasus III mass spectrometry system (LECO). Mass spectral recording was set to 20 scans/s. Transfer line and ion source temperatures were set to 250 °C. The monitored mass range was *m/z* 70–600 amu. This range was extended to *m/z* 45–1000 amu for recording mass spectral tag (MST) information, namely RI and full mass spectrum, of reference compounds. Pipetting steps, automated chemical derivatisation and timed in-line injection into the GC–MS system were performed using a CTC Combi PAL autosampler and PAL cycle composer software version 1.5.0 (CTC Analytics AG, Zwingen, Switzerland).

2.4.2. Variant 2 (5PDM_RTX5_9_TOF)

Variant 2 differed only by choice of an alternative 5PDM capillary column type with equal dimensions, namely a 0.25 μm, 30 m × 0.25 mm RTX-5Sil MS with 10 m integrated guard column (Restek GmbH, Bad Homburg, Germany). The mass range was set to *m/z* 70–600 amu.

2.4.3. Variant 3 (5PDM_RTX5_15_TOF)

Like variant 2, variant 3 [18] had a 0.25 μm, 30 m × 0.25 mm RTX-5Sil MS capillary column with 10 m integrated guard column (Restek), but the temperature programming was altered to 2 min isothermal period at 80 °C, 15 °C/min ramp to 350 °C and 2 min at final temperature. Injection was splitless at 230 °C with a 2 min 110 psi pressure pulse at constant 1 mL/min flow rate. TOF-detection was performed using a Pegasus II mass spectrometry system (LECO). Mass spectral recording was adjusted to 6 scans/s and *m/z* 70–600 amu. The ion source temperature and transfer line were set to 200 °C and 250 °C.

2.4.4. Variant 4 (5PDM_DB5_40_TOF)

The method variant 4 [29] was a fast GC–TOF–MS application on a 6890N gas chromatograph (Agilent Technologies) hyphenated to a Pegasus mass spectrometry system (LECO). A 5PDM DB5-MS fused silica capillary column with 0.18 μm, 10 m × 0.18 mm dimensions (J&W Scientific, Folsom, CA, USA) was operated 2 min at 70 °C followed by a 40 °C/min ramp to 320 °C and a 1 min heating at final temperature. Injection was 1 μL at 270 °C with 1 min purge time at 20 mL/min purge flow. The transfer line and the ion source were set to 250 °C and 200 °C, respectively. The scan rate was 30 scans/s at *m/z* 50–800 amu.

2.4.5. Variant 5 (5PDM_VF5_6_Q)

A Trace GC ultra gas chromatograph with an AS 3000 auto sampler and a DSQ quadrupole-type mass spectrometer (ThermoFinnigan, San Jose, CA, USA) was used by variant 5 [30]. The sample was injected at 230 °C and separated on a 5PDM-type VF-5 ms 0.25 μm, 30 m × 0.25 mm fused silica capillary column (Varian), with helium at a flow rate of 1 mL/min. Temperature programming was 1 min isothermal at 70 °C, followed by 1 °C/min to 76 °C and 6 °C/min to 330 °C with 10 min final heating at 330 °C. Mass spectra were monitored with *m/z* 70–600 amu and 2 scans/s. The transfer line was set to 280 °C and the ion source to 250 °C.

2.4.6. Variant 6 (5PDM_DB5_6_Q)

Variant 6 [31] utilised a Trace gas chromatograph mounted with an AS 2000 auto sampler and a Trace mass spectrometer (ThermoFinnigan). Gas chromatography was performed with a 5PDM-type capillary column, namely a DB5-MS fused silica capillary column with 0.25 μm, 30 m × 0.25 mm dimensions (J&W Scientific) and helium carrier gas at 1 mL/min. Temperature programming was 1 min isothermal at 70 °C, followed by 1 °C/min to 76 °C and 6 °C/min to 325 °C with 10 min heating at 325 °C. The ion source temperature was adjusted to 220 °C. Mass spectra were recorded at 2 scans/s with *m/z* adjusted to 35–573 amu.

2.4.7. Variant 7 (5PDM_RTX5_5_Q)

Variant 7 [32] was performed using GC 8000 gas chromatograph coupled to a Voyager quadrupole-type mass spectrometer and an AS 2000 auto sampler (ThermoFinnigan). Gas chromatography was performed on a 0.25 μm, 30 m × 0.25 mm RTX-5Sil MS capillary column with 10 m integrated guard column (Restek) with 5 min isothermal period at 70 °C, a 5 °C/min temperature ramp to 320 °C and 1 min final heating. Sample injection was splitless at 230 °C and 1 mL/min helium carrier flow. The interface to the mass spectrometer was set to 250 °C and the ion source adjusted to 200 °C. The monitored mass range was set to *m/z* 40–600 amu. Mass spectra were recorded at 1.67 scans/s.

2.4.8. Variant 8 (5PDM_Eq5_3_TRAP)

The variant 8 [33] used an ion trap-type mass spectrometer, namely a PolarisQ ion trap mass spectrometer equipped with a Trace GC gas chromatograph and an AS2000 auto sampler (ThermoFinnigan). Splitless injection at 250 °C was performed with constant flow settings, 1 mL/min helium. A 5-PDM type capillary column was mounted, the Equity-5 column, 0.25 μm, 30 m × 0.25 mm (Supelco, Bellfonte, CA, USA). Chromatography settings were 3 min at 80 °C and 3 °C/min to 300 °C. The mass spectral acquisition rate was 2 scans/s with a range of *m/z* 50–550 amu. Transfer line and ion source temperatures were set to 250 °C and 200 °C, respectively.

2.4.9. Variant 9 (35PDM_MDN35_15_TOF)

Variant 9 represents the only chromatographic system of this study with capillary column polarity changed to a 35PDM-type [27]. The GC–TOF–MS system and basic settings were as described of variant 1. Except, the chosen capillary column, MDN-35, 0.25 μm, 30 m × 0.32 mm (Sigma–Aldrich), was operated at constant 2 mL/min helium flow starting with 2 min at 80 °C, heat ramping 15 °C/min to 330 °C and completing the cycle with 6 min at 330 °C.

2.4.10. Retention time standardisation

Method variants 1 and 9 used a combination of *n*-alkane mixture and FAME mixture for retention time standardisation and estimation of accuracy of prediction and precision of measurement and calculation (cf. Section 2.1). All other variants employed the above *n*-alkane mixture with the following variations. Variants 3–5 and 9 omitted *n*-decane because of chromatographic limitations. Variants 3, 5, and 6 lacked *n*-octadecane, whereas variant 6 had the complete set of *n*-alkanes ranging from C_{12} to C_{25}. The RIs of analytes, which were not bracketed by two retention markers, were extrapolated. Regression procedures based on all available marker compounds were applied without forcing an intercept. Alternatively calculations were performed based on the two nearest neighbours, for example the interpolation procedure according to algorithm proposed by van den Dool and Kratz [34]. Precision and accuracy were expressed in terms of standard deviation using either *n*-alkane based RI units or percent of the average and percent of expected, respectively.

2.5. Retention time retrieval, calculations and statistical procedures

The retention times from method variants 4–8 were retrieved at local chromatographic peak apices. Compound identity was manually confirmed by mass spectral match. Variants 1, 2, 3 and 9 were automatically deconvoluted [27] and mass spectra matched to a reference library through ChromaTof software (LECO). Peak lists of non-normalised mass spectra were exported and processed by TagFinder software [24]. Retention times were retrieved from these peak list files using the retention index calculation tool of the TagFinder software searching for retention times at local abundance maxima of compound characteristic mass fragments, such as *m/z* 71, 85, 99, 113 amu of *n*-alkanes and *m/z* 74, 87, 101, 143 amu of FAMEs and respective molecular masses. TagFinder has only van den Dool calculation of RIs implemented.

A Microsoft SQL Server 2005® was used as the relational database backend for storage and management of the mass spectral and chromatographic retention library information. Algorithms for RI processing were implemented using the Common Language Runtime (CLR.net), the C# programming language and Microsoft Visual Studio 2005®. Retention indices of analytes were computed using user-defined functions (UDF) of the database and T-SQL to access retention times of analytes and both the retention times and retention index definitions of the *n*-alkane or FAME marker compounds. Exploratory data visualisation was performed using Microsoft Excel software.

3. Results and discussion

3.1. Precision of empirical RI determination

Information on the empirical precision and accuracy of RI determination is prerequisite for the evaluation of RI projection methods, which aim to utilise existing RI libraries, such as provided by GMD, http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html, for the transfer of retention properties between chromatographic method variants. Ultimately, the threshold settings for compound matching will depend both on the achievable exactness of determination and projection.

3.1.1. Retention time drift

Retention time drift is one of the main obstacles to the utilisation of chromatographic compound properties for chemical identification purposes. Variations of capillary column length and artefacts of injection timing (Fig. 1) or slight changes in flow and pressure settings may strongly affect observable retention times. Moreover, capillary columns for gas chromatography have a limited life time which is limited by a slow continuous retention drift caused by gradual loss and modification of the stationary phase. These altered column properties may speed up chromatography significantly in

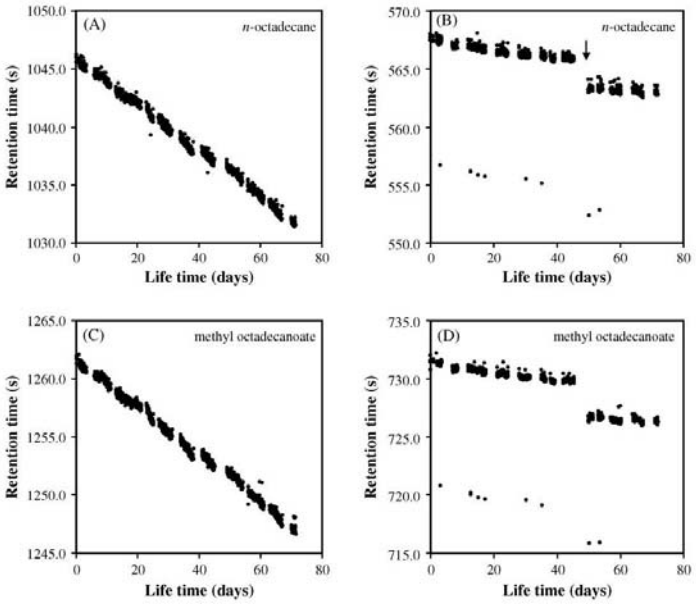


Fig. 1. Retention time drift of *n*-octadecane and methyl stearate under routine operating conditions of metabolite profiling analyses. Exemplary sequences of ~1340 chromatograms from variant 1 (A and C) and variant 9 (B and D) are displayed. Variants differ in column polarity, namely 5%-phenyl-95%-dimethylpolysiloxane versus 35%-phenyl-65%-dimethylpolysiloxane equivalent stationary phases, and temperature ramp, 9 °C/min versus 15 °C/min. Note that the shortening of capillary columns (arrow) may become necessary during routine maintenance of both method variants. The retention time outliers (cf. B and D) are artefacts caused by early injection. This artefact is not specific of variant 9. Both columns were conditioned by reagent injections prior to this selection of chromatographic runs. Total life time of both capillary columns may exceed the shown period.

the course of 1000 sample injections (Fig. 1). This drift depends on temperature ramping and column stability among other factors.

In view of these factors, we decided to apply the traditional concept of chemical retention time standardisation instead of relying on mathematical chromatography alignment procedures. In this study we used an *n*-alkane based RI system, where RI is defined as the number of carbon atoms multiplied by 100, and measurement performed by internal retention standardisation of each single chromatographic run. We selected variants 1 and 9 for investigations of the precision and accuracy of empirical RI determination, because the mass spectral scanning rate of these variants had sufficiently high resolution, namely an average of 0.086 RI units/scan and 0.136 RI units/scan, respectively, as determined through the distance by number of scans between dodecane and hexatriacontane peak apices. In comparison, other chromatographic variants constituting GMD, such as variant 7, have considerably lower chromatographic resolution, e.g. 0.679 RI units/scan.

3.1.2. *Precision of retention index calculations from low complexity profiles*

In the following we will demonstrate the influence of calculation methods and sample matrix on RI measurements. We partitioned the chromatographic runs of our study (Fig. 1) into sets comprising low complexity profiles, profiles containing a matrix of defined authenticated reference substances, DRM-mix, and profiles containing either intracellular metabolites of a microbial matrix, DRM-yeast, or a highly complex plant reference sample, DRM-rice. Low complexity profiles represented non-sample control runs or included a single reference substance, and thus typical mapping experiments to obtain reference RI properties. Retention times of spiked *n*-alkanes and FAMEs were retrieved from each chromatogram and RIs of FAMEs calculated using either interpolation methods, namely van den Dool [34], Kováts [35] and spline algorithms, or polynomial regression models using 1st–7th order and exponential fitting (Table 2). In the following, the precision of empirical RI determination is expressed in terms of RI standard deviation with independent replication >40 and accuracy estimated by difference of RI determined in the presence of a complex biological matrix compared to low complexity samples.

Precision of alkane RI determinations approximated chromatographic resolution in low complexity samples, when regression with increasing order was employed. However, average precision of FAME RIs remained limited to 0.22–0.33 RI units, using Van den Dool and spline interpolation or 3rd–5th order regression; exponential fitting was found to be non-optimal. These observations were made for both chromatography variants (Table 2).

3.1.3. *Precision of retention index calculations from high complexity profiles*

We selected sample types with increasing chemical complexity, namely DRM-mix < DRM-yeast < DRM-rice, to estimate the impact of matrix composition on RI determinations. The DRM-mix of 28 reference substances (Table 2) and varying mixtures of 20–25 substances (data not shown) did not affect RI determinations. However, both biological matrices had a negative effect. Average RI (S.D.) may increase to 0.75–1.11, depending on sample type and chromatography variant (Table 2).

Table 2
Influence of matrix composition on the precision of RI determination[a]

Method of calculation	Precision of RI determination (average standard deviation)							
	Average of C_{10}–C_{36} n-alkanes				Average of C_8–C_{30} Fatty Acid Methylesters			
	Low complexity	DRM-mix	DRM-yeast	DRM-rice	Low complexity	DRM-mix	DRM-yeast	DRM-rice
Chromatography variant 1								
1st Order polynomial regression	0.1788	0.1734	0.1839	2.6142	0.2274	0.2622	0.2633	3.5782
2nd Order polynomial regression	0.1525	0.1354	0.1546	0.8669	0.2342	0.2585	0.2633	1.5822
3rd Order polynomial regression	**0.1060**	**0.0957**	**0.1056**	**0.5721**	**0.2225**	**0.2494**	**0.2520**	**1.3407**
4th Order polynomial regression	0.0662	0.0469	0.0618	0.3339	0.2209	0.2487	0.2631	1.1161
5th Order polynomial regression	0.0407	0.0295	0.0410	0.1390	0.2234	0.2508	0.2640	1.1035
6th Order polynomial regression	0.0149	0.0101	0.0135	0.0449	0.2222	0.2475	0.2687	1.8134
Exponential regression	0.4098	0.4050	0.4066	5.3849	0.3582	0.4114	0.3576	5.1470
Spline interpolation					0.2326	0.2553	0.2689	1.1139
Kovàts interpolation					0.2252	0.2402	0.2640	1.1555
Van den Dool interpolation					**0.2128**	**0.2326**	**0.2510**	**1.1070**
Chromatography variant 9								
1st Order polynomial regression	0.1997	0.1551	0.3390	0.3958	0.2425	0.2499	0.4866	0.7146
2nd Order polynomial regression	0.1408	0.1054	0.2744	0.3285	0.2425	0.2465	0.4901	0.7496
3rd Order polynomial regression	**0.0975**	**0.0782**	**0.2446**	**0.2957**	**0.2513**	**0.2416**	**0.4892**	**0.7296**
4th Order polynomial regression	0.0651	0.0587	0.2059	0.2341	0.2674	0.2493	0.5332	0.6965
5th Order polynomial regression	0.0305	0.0293	0.1776	0.2004	0.2921	0.2641	0.5750	0.7180
6th Order polynomial regression	< 0.0001	0.0001	0.0002	0.0002	0.5288	0.4584	2.2751	2.5882
Exponential regression	0.3328	0.3364	0.4787	0.5515	0.3143	0.3496	0.4936	0.7779
Spline interpolation					0.2764	0.2515	0.5618	0.7995
Kovàts interpolation					0.5809	0.2769	0.5893	0.8377
Van den Dool interpolation					**0.3278**	**0.2527**	**0.5269**	**0.7490**

[a] Regression and interpolation methods were applied to calculate retention indices of n-alkanes and fatty acid methyl esters spiked into routine GC–TOF–MS metabolite profiles. Two chromatography variants, 1 or 9 (cf. Fig. 1), are compared. Samples had either low complexity, namely single reference substances and non-sample controls, or comprised complex defined reference material (DRM) of 28 reference substances (DRM-mix, $n = 45$), intracellular extracts of yeast (DRM-yeast, $n = 45$) and of rice (DRM-rice, $n = 41$). Van den Dool interpolation and 3rd order regression are highlighted by bold font.

All interpolation methods, namely spline, Kováts and van den Dool, and most regression algorithms were equally sensitive to these matrix effects. However, exponential fit, 1st, 2nd, 6th and higher order regression models did not properly reflect the impact of matrix on retention shifts.

In the following we selected van den Dool interpolation to investigate the source of reduced RI precision (Fig. 2). The strongest matrix effects were observed in early parts of the temperature programming of both investigated chromatography variants, namely the C_8 ME–C_9 ME region (Fig. 2A). The increased RI (S.D.) coincided with delayed retention of C_8 ME–C_9 ME. Thus, early eluting FAMEs had the strongest impact on overall RI accuracy in the presence of biological matrix with average $|RI_{(DRM\text{-}rice)} - RI_{(low\ complexity)}|$ equal to 1.0 (variant 1) and 0.5 (variant 9) RI units, respectively. The 3rd order regression algorithm was tested in parallel and exhibited highly similar results (data not shown). Therefore, we concluded that van den Dool interpolation and 3rd order regression are equivalent calculation approaches with respect to RI precision and accuracy.

3.1.4. Comparison of variant 1 and 9

The comparison of chromatography variants 1 and 9 demonstrated enhanced retention time stability of variant 9 (Fig. 1). The reduced retention drift appears to propagate into slightly enhanced RI precision (Fig. 2A) and accuracy (Fig. 2B). The cause of the improved retention behaviour of variant 9 was not further investigated and was deemed beyond the scope of this study. Both, the reduced duration of exposure to high temperatures per analysis cycle and possibly the altered stability of the capillary column may contribute. Moreover, the impact of matrix on RI performance may change with the biological object under investigation. For example, variant 9 performed better in the presence of DRM-rice, whereas variant 1 appeared to exhibit improved results with DRM-yeast.

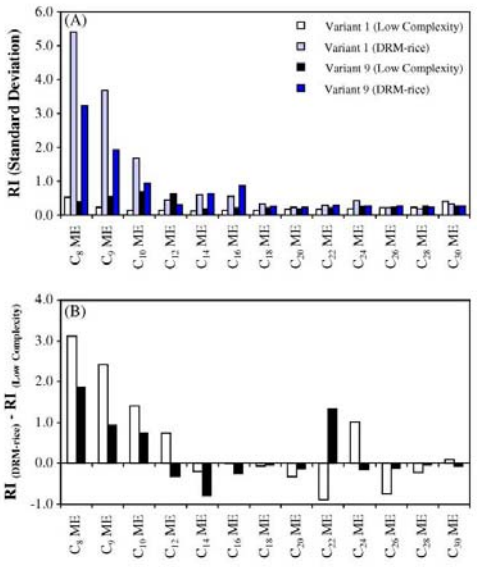


Fig. 2. The matrix effect negatively affecting RI precision and accuracy of methyl esters (ME) is dependent on chromatographic region. (A) Precision of RI determination was calculated as the standard deviation of RI ($n > 40$). (B) Accuracy of RI determination was estimated by comparison of RI measured in the presence of a complex biological matrix compared to a low complexity chemical background; the average $|RI_{(DRM\text{-}rice)} - RI_{(low\ complexity)}|$ was 1.0 (variant 1) and 0.5 (variant 9) RI units. Note that the strongest matrix effects occur at the start of the temperature program, e.g. C_8 ME and C_9 ME.

Table 3
Transfer of retention index (RI) properties between chromatography variants of the Golm Metabolome Database (GMD)[a]

	Variant 1	Variant 2	Variant 3	Variant 4	Variant 5	Variant 6	Variant 7	Variant 8
Column brand	VF5	RTX5	RTX5	DB5	VF5	DB5	RTX5	Eq5
Temperature ramp (°C/min)	9	9	15	40	6	6	5	3
Gas chromatograph	6890N	6890N	6890N	6890N	Trace GC ultra	Trace GC	GC8000	Trace GC
Aquisition rate (scans/s)	20	20	6	30	2	2	1.67	2
Mode of detection	TOF	TOF	TOF	TOF	Q	Q	Q	TRAP
A	Number of paired analytes							
Variant 1	488	348	274	179	157	175	318	65
Variant 2		931	623	209	226	244	518	96
Variant 3			964	184	206	192	437	93
Variant 4				299	151	127	197	71
Variant 5					264	154	224	77
Variant 6						324	190	82
Variant 7							961	96
Variant 8								103
B	Correlation coefficient (r^2)							
Variant 1	–	0.99984	0.99971	0.99945	0.99914	0.99951	0.99926	0.99956
Variant 2	0.99983	–	0.99979	0.99951	0.99910	0.99961	0.99932	0.99962
Variant 3	0.99971	0.99979	–	0.99950	0.99973	0.99977	0.99977	0.99945
Variant 4	0.99938	0.99945	0.99943	–	0.99880	0.99983	0.99886	0.99976
Variant 5	0.99916	0.99913	0.99974	0.99904	–	0.99987	0.99974	0.99962
Variant 6	0.99951	0.99961	0.99977	0.99983	0.99987	–	0.99981	0.99975
Variant 7	0.99925	0.99931	0.99977	0.99910	0.99974	0.99981	–	0.99978
Variant 8	0.99956	0.99962	0.99945	0.99976	0.99963	0.99975	0.99978	–
C	Standard deviation ($RI_{predicted} - RI_{determined}$)							
Variant 1	–	7.42	9.21	12.71	16.05	9.54	16.04	9.55
Variant 2	7.50	–	8.00	12.01	17.13	9.33	15.00	8.20
Variant 3	9.30	8.10	–	12.12	9.18	7.83	9.29	9.58
Variant 4	13.64	12.89	13.08	–	19.81	5.69	19.58	7.36
Variant 5	15.93	16.82	9.14	17.56	–	5.11	9.36	9.22
Variant 6	9.56	9.37	7.73	5.66	5.11	–	6.75	6.83
Variant 7	16.19	15.04	9.27	17.17	9.28	6.70	–	6.80
Variant 8	9.07	7.82	9.10	6.98	8.72	6.50	6.53	–
D	Standard deviation ($RI_{predicted} - RI_{determined}$ [% of $RI_{determined}$])							
Variant 1	–	0.38	0.42	0.52	0.59	0.55	0.64	0.52
Variant 2	0.39	–	0.42	0.47	0.61	0.64	0.60	0.44
Variant 3	0.43	0.44	–	0.54	0.46	0.51	0.43	0.57
Variant 4	0.53	0.48	0.54	–	0.66	0.35	0.69	0.40
Variant 5	0.59	0.62	0.45	0.62	–	0.29	0.40	0.46
Variant 6	0.56	0.67	0.50	0.35	0.29	–	0.37	0.38
Variant 7	0.64	0.60	0.42	0.64	0.40	0.36	–	0.36
Variant 8	0.50	0.43	0.55	0.38	0.44	0.37	0.35	–

[a] All included method variants were based on 5%-phenyl-95%-dimethylpolysiloxane or equivalent stationary phases and were operated at 1 mL/min constant helium flow. Column brand, temperature programming and mass spectral detection varied as indicated. (A) Number of paired analytes, which were used for 3rd order polynomial regression, (B) regression coefficients, r^2, (C) accuracy of prediction as characterised by standard deviation of residual errors, $RI_{predicted} - RI_{determined}$, and (D) accuracy of prediction as characterised by standard deviation of residual errors expressed as percent of $RI_{determined}$. Note that resulting matrices B and C are not exactly symmetrical; horizontal variants were used to predict RIs of the variants listed vertically.

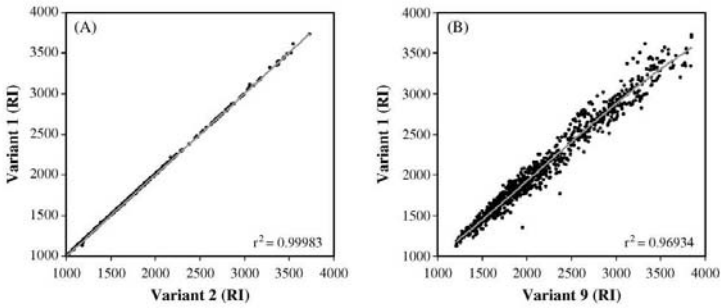


Fig. 3. Comparison of correlation of RI systems using 3rd order polynomial regression. (A) Variant 2 and variant 1 have equivalent 5%-phenyl-95%-dimethylpolysiloxane stationary phases. (B) In contrast, variant 9 utilises a 35%-phenyl-65%-dimethylpolysiloxane stationary phase. The fitted functions and correlation coefficients, r^2, are shown.

3.1.5. Check of biochemically relevant analytes

Precision of RI determination from our study was similar to the ~1 RI unit precision reported of a set of 250 volatile analytes [36]. We used the above DRM-mix (Table 1) to estimate, if results of RI precision, which were based on FAME, may be transferred to those compound classes which are relevant for routine metabolite profiling experiments. In the following, the chosen exemplary analytes are listed with RI (S.D.) of low complexity samples and – in square brackets – the respective precisions determined in the presence of DRM-yeast followed by DRM-rice. We analysed RI (S.D.) of citric acid (4TMS), 0.37 [0.43; 1.24], valine (2TMS), 0.30 [0.42; 2.51], glycine (3TMS), 0.20 [0.28; 1.94], ribitol (5TMS), 0.54 [0.53; 0.83], as well as glucose (1MEOX) (4TMS), 0.63 [4.00; 0.85]. In conclusion we found the results obtained with our FAME mixture to be representative. Precisions were influenced by matrix rather than by nature of chemical compound.

3.2. Transfer of RI properties between chromatography variants

The GMD, http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html, collection of mass spectral tags, a combination of mass spectral and RI library, was initially compiled with the intention of minimum redundancy and maximum complementation of diverse reference substances. For shared use of these data, the means of transferring RI information between the chromatography variants constituting GMD became essential. With increasing numbers of entries a substantial portion of redundancy became available (Table 3A). Sets of 65–623 paired analytes between chromatography variants allowed statistically sound investigation of regression models for RI prediction and estimation of residual error.

A correlation analysis of the retention index systems, which were measured using 5%-phenyl-95%-dimethylpolysiloxane or equivalent stationary phases, demonstrated high apparently linear correlation (Fig. 3A). This high linearity strongly contrasted with the expected low linear fit, when the RI system of variant 1 was fitted to variant 9, which utilises a 35%-phenyl-65%-dimethylpolysiloxane stationary phase (Fig. 3B). A detailed analysis of polynomial regression models applied among variants 1–8 revealed that in most cases 3rd order polynomial regression was sufficient to obtain a small increase of fit. Correlation coefficients, r^2, were improved at and beyond 3rd order regression, as was the standard deviation of the residual error, which was determined by $RI_{predicted} - RI_{determined}$ (Fig. 4). Two factors contributed to the magnitude of this residual error. (1) The residual error appeared to be a function of RI and thus may be proportional to the boiling point of analytes (Fig. 5B). (2) Single analytes may exhibit abnormally high deviations (Fig. 5). This abnormal behaviour could not be linked to a single type of analyte or the influence of specific chemical groups as determinant chemical features (data not shown). However, the amount of analyte was demonstrated earlier to have an impact on RI behaviour [19]. In this study we decided to keep the amount of substance constant for the purpose of RI mapping (cf. Sections 2.1 and 2.2) and attribute abnormally high deviations to non-documented quantitative experimental errors of previous method variants. Erroneous analyte assignments had been eliminated earlier.

Following the principle of making the minimal required number of assumptions we decided for a 3rd order regression model and investigated all possible pair wise predictions between chromatography variants (Table 3B–D). Regression coefficients ranged from 0.99880 to 0.99987. The standard deviation of residual errors varied from 5.1 to 19.8 RI units, equivalent to 0.29–0.69%. The margin of error was, thus, similar to the robustness reported of non-derivatised volatile analytes [36]. In comparison standard deviation of residual error was 108 (4.44%) RI units, when the 35-PDM variant 9 was used to predict variant 1 (Fig. 3B). The accuracy of RI transfer between chromatography variants appeared not to be subject to general systematic factors, except that variants with a shallow temperature ramp, namely variants 5–8, appeared to match better among each other. A similar observation was made with variants 1–3. In contrast to the apparent trend, the fast GC application of variant 4 had, however, best agreement with variants 6 and 8.

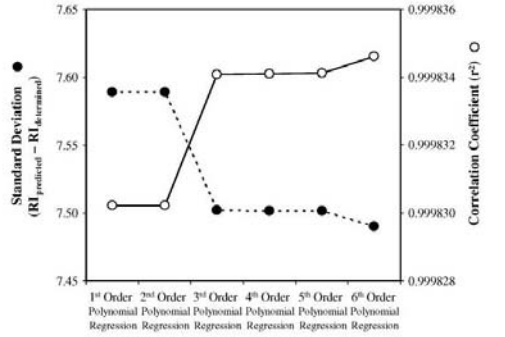


Fig. 4. Correlation coefficients, r^2 (open circle), and accuracy of prediction as determined by the standard deviation of the residual error (closed circle). Residual error of paired analytes was determined by $RI_{predicted} - RI_{determined}$. The projection of variant 2 onto variant 1 was subjected to permutation of 1st–6th order polynomial regression. 3rd order regression was found to represent the minimum required assumption for optimal prediction.

Finally we investigated, if grouping by chemical nature of analytes may improve accuracy of RI transfer between chromatography

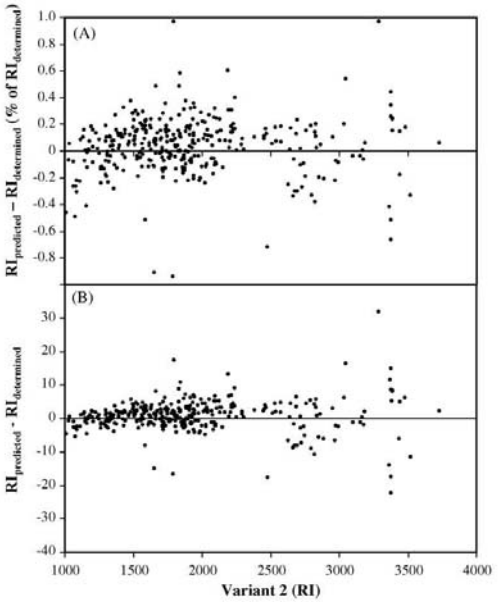


Fig. 5. Residual error of 3rd order polynomial regression using RI information of variant 2 to predict variant 1. (A) Percent of RI deviation $(RI_{predicted} - RI_{determined}) \times RI_{determined}^{-1} \times 100$, overall standard deviation, 0.39% ($n = 348$). (B) absolute RI deviation, $RI_{predicted} - RI_{determined}$, overall standard deviation 7.50 RI units (cf. Fig. 3A).

variants. This analysis was based on the report of Stein and co-authors on the use of the so-called group contributions to estimate Kováts RIs [21]. In the following we report examples taken from the projection of variant 2 onto variant 1 which had overall 7.42 standard deviation (S.D.) residual error, equivalent to 0.39% relative standard deviation and $r^2 = 0.99984$ with $n = 348$ paired analytes (Fig. 5). The subset of 29 hydroxy-, di- and tricarboxylic acids had 0.78 (S.D.) and 0.999989 (r^2). Moreover, a combination of all 37 paired sugars resulted in 3.74 (S.D.) with 0.999968 (r^2) and the set of 12 polyols and primary alcohols had 1.57 (S.D.) with $r^2 = 0.999985$. On the other hand a set of 12 compounds with purine, pyrimidine and indole *N*-heterocycles exhibited no improvement, e.g. 7.14 (S.D.) and $r^2 = 0.999285$.

4. Conclusions

We demonstrated that equal precision of RI determination could be obtained by previously reported interpolation methods [34,35] as well as spline and 3rd–5th order polynomial regression procedures. These findings held true in the presence of defined matrices and highly complex extracts from yeast cells and rice plants. For RI calculations within GMD [22,23] and the TagFinder software [24] we implemented the conventional van den Dool algorithm to best agree with earlier reports.

For transfer of RI information between chromatography variants using identical polarity of the stationary phase we selected a 3rd order regression model and implemented a respective projection procedure within GMD. This transfer procedure will provide mass spectra from GMD with RI predictions of those compounds which do not have experimentally verified variant RIs.

Moreover, we clearly demonstrated three possible levels of selecting RI thresholds. (1) In the presence of low complexity samples a threshold of 0.25 RI units may be applicable. This threshold, however, strictly applies only to controlled amounts of standards and analytes (cf. Section 2.1) (2) In the presence of highly complex samples the threshold must be set at least one order of magnitude higher. In the early chromatographic region analytes may exceed the respective threshold of approximately 2–3 RI units. (3) When projections from other chromatography variants are used the thresholds may be inferred from the standard deviations of residual errors (Table 3). As demonstrated by Fig. 5 thresholds may best be set as percent error of the expected absolute RI, for example to 0.5–1.0% (cf. Table 3D and Fig. 5).

In conclusion, accuracy of RI prediction may be much improved compared to earlier reports, when strictly equivalent stationary phases are exclusively considered. However, the estimated thresholds remain in part too broad for an unambiguous identification of isomers, especially in the presence of complex biological matrix. Therefore, we recommend for routine profiling analyses the co-analysis of defined mixtures of reference substances with each single GC–MS metabolite profiling experiment. These reference mixtures should be adjusted in quantitative and qualitative composition to the respective biological matrix under investigations. Mixtures may comprise (1) sets of authenticated reference substances which should cover the range of expected metabolite classes and (2) should contain selected isomers of those compound classes which cannot be distinguished by mass spectrometry alone.

Acknowledgements

This work was supported by the Max Planck Society, the Quant-Pro program of the Bundesministerium für Bildung und Forschung (BMBF), sub-project "InnOx–Innovative diagnostic tools to optimise potato breeding: Systemic analysis of cellular processes and their relation to plant internal oxygen concentrations", FKZ 0313813A, and the European META-PHOR project, FOOD-CT-2006-036220. The authors acknowledge the long standing support and encouragement by Prof. L. Willmitzer, Max Planck Institute of Molecular Plant Physiology (MPI-MP), Am Muehlenberg 1, D-14476 Potsdam-Golm, Germany.

References

[1] J.K. Nicholson, J.C. Lindon, E. Holmes, Xenobiotica 29 (1999) 1181.
[2] R.N. Trethewey, A.J. Krotzky, L. Willmitzer, Curr. Opin. Plant Biol. 2 (1999) 83.
[3] L.W. Sumner, P. Mendes, R.A. Dixon, Phytochemistry 62 (2003) 817.
[4] R.J. Bino, R.D. Hall, O. Fiehn, J. Kopka, K. Saito, J. Draper, B.J. Nikolau, P. Mendes, U. Roessner-Tunali, M.H. Beale, R.N. Trethewey, B.M. Lange, E.S. Wurtele, L.W. Sumner, Plant Sci. 9 (2004) 418.
[5] G. Stephanopoulos, H. Alper, J. Moxley, Nat. Biotechnol. 22 (2004) 1261.
[6] J. Nielsen, S.G. Oliver, Trends Biotechnol. 23 (2005) 544.
[7] H. Jenkins, N. Hardy, M. Beckmann, J. Draper, A.R. Smith, J. Taylor, O. Fiehn, R. Goodacre, R.J. Bino, R.D. Hall, J. Kopka, G.A. Lane, B.M. Lange, J.R. Liu, P. Mendes, B.J. Nikolau, S.G. Oliver, N.W. Paton, S. Rhee, U. Roessner-Tunali, K. Saito, J. Smedsgaard, L.W. Sumner, T. Wang, S. Walsh, E.S. Wurtele, D.B. Kell, Nat. Biotechnol. 22 (2004) 1601.
[8] J.C. Lindon, H.C. Keun, T.M.D. Ebbels, J.M.T. Pearce, E. Holmes, J.K. Nicholson, Pharmacogenomics 6 (2005) (2005) 691.
[9] A.L. Castle, O. Fiehn, R. Kaddurah-Daouk, J.C. Lindon, Brief. Bioinform. 7 (2006) 159.
[10] O. Fiehn, B. Kristal, B. Van Ommen, L.W. Sumner, S.A. Sansone, C. Taylor, N. Hardy, R. Kaddurah-Daouk, Omics J. Intergrat. Biol. 10 (2006) 158.
[11] O. Fiehn, D. Robertson, J. Griffin, M. van der Werf, B.J. Nikolau, N. Morrison, L.W. Sumner, R. Goodacre, N.W. Hardy, C. Taylor, J. Fostel, B. Kristal, R. Kaddurah-Daouk, P. Mendes, B. van Ommen, J.C. Lindon, S.A. Sansone, Metabolomics 3 (2007) 175.
[12] L.W. Sumner, A. Amberg, D. Barrett, M.H. Beale, R. Beger, C.A. Daykin, T.W.M. Fan, O. Fiehn, R. Goodacre, J.L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A.N. Lane, J.C. Lindon, P. Marriott, A.W. Nicholls, M.D. Reily, J.J. Thaden, M.R. Viant, Metabolomics 3 (2007) 211.
[13] O. Fiehn, J. Kopka, P. Dörmann, T. Altmann, R. Trethewey, L. Willmitzer, Nat. Biotechnol. 18 (2000) 1157.
[14] U. Roessner, C. Wagner, J. Kopka, R.N. Trethewey, L. Willmitzer, Plant J. 23 (2000) 131.
[15] P. Ausloos, C.L. Clifton, S.G. Lias, A.I. Mikaya, S.E. Stein, D.V. Tchekhovskoi, O.D. Sparkman, V. Zaikin, D. Zhu, J. Am. Soc. Mass Spectrom. 10 (1999) 287.
[16] J.M. Halket, D. Waterman, A.M. Przyborowska, R.K.P. Patel, P.D. Fraser, P.M. Bramley, J. Exp. Bot. 56 (2005) 219.
[17] S.E. Stein, J. Am. Soc. Mass Spectrom. 10 (1999) 770.
[18] C. Wagner, M. Sefkow, J. Kopka, Phytochemistry 62 (2003) 887.
[19] J. Kopka, J. Biotechnol. 124 (2006) 312.
[20] V.I. Babushok, P.J. Linstrom, J.J. Reed, I.G. Zenkevich, R.L. Brown, W.G. Mallard, S.E. Stein, J. Chromatogr. A 1157 (2007) 414.
[21] S.E. Stein, V.I. Babushok, R.L. Brown, P.J. Linstrom, J. Chem. Inf. Model. 47 (2007) 975.
[22] N. Schauer, D. Steinhauser, S. Strelkov, D. Schomburg, G. Allison, T. Moritz, K. Lundgren, U. Roessner-Tunali, M.G. Forbes, L. Willmitzer, A.R. Fernie, J. Kopka, FEBS Lett. 579 (2005) 1332.
[23] J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmueller, P. Doermann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A.R. Fernie, D. Steinhauser, Bioinformatics 21 (2005) 1635.
[24] A. Luedemann, K. Strassburg, A. Erban, J. Kopka, Bioinformatics 24 (2008) 732.
[25] C.T. Peng, Z.C. Yang, D. Maltby, J. Chromatogr. 586 (1991) 113.
[26] A. Erban, N. Schauer, A.R. Fernie, J. Kopka, in: W. Weckwerth (Ed.), Metabolomics: Methods and Protocols, Humana Press, Totowa, 2007, pp. 19–38.
[27] J. Lisec, N. Schauer, J. Kopka, L. Wilmitzer, A. Fernie, Nat. Protoc. 1 (2006) 387.
[28] X. Yang, V. Romheld, H. Marschner, Plant Soil 164 (1994) 1.
[29] J. Gullberg, P. Jonsson, A. Nordström, M. Sjöström, T. Moritz, Anal. Biochem. 331 (2004) 283.
[30] U. Roessner, J. Patterson, M.G. Forbes, G. Fincher, P. Langridge, A. Bacic, Plant Physiol. 142 (2006) 1087.
[31] S. Strelkov, M. von Elstermann, D. Schomburg, Biol. Chem. 385 (2004) 853.
[32] G. Colebatch, G.G. Desbrosses, T. Ott, L. Krusell, O. Montanari, S. Kloska, J. Kopka, M.K. Udvardi, Plant J. 39 (2004) 487.
[33] A. Barsch, T. Patschkowski, K. Niehaus, Funct. Integr. Genomics 4 (2004) 219.
[34] H. van den Dool, P.D. Kratz, J. Chromatogr. 11 (1963) 463.
[35] E. Kováts, Helv. Chim. Acta 41 (1958) 1915.
[36] F. Bianchi, M. Careri, A. Mangia, M. Musci, J. Sep. Sci. 30 (2007) 563.

Appendix C: Supporting Software Development and Statistical Datamining of Transcript Profiles

[15] Steinhauser D, Junker BH, Luedemann A, Selbig J, **Kopka J** (2004) Hypothesis-driven approach to predict transcriptional units from gene expression data. *Bioinformatics* 20 (12): 1928-1939
(http://dx.doi.org/10.1093/bioinformatics/bth182)
(http://bioinformatics.oxfordjournals.org/cgi/content/abstract/20/12/1928)
Free article (http://bioinformatics.oxfordjournals.org/cgi/reprint/20/12/1928)

[16] Steinhauser D, Usadel B, Luedemann L, Thimm O, **Kopka J** (2004) CSB.DB: A comprehensive systems-biology database. *Bioinformatics* 20 (18): 3647-3651
(http://dx.doi.org/10.1093/bioinformatics/bth398)
(http://bioinformatics.oxfordjournals.org/cgi/content/abstract/20/18/3647)
Free article (http://bioinformatics.oxfordjournals.org/cgi/reprint/20/18/3647)

Project leadership of investigations aimed at the analysis of the potential and limitations of correlation algorithms for large profiling data sets. The project utilized compendia of transcript profiles and provided an internet-based application which allows retrieval of correlated genes from sets of predefined transcript profiles. Among other tools, intersection correlation analysis between multiple genes and gene function enrichment analysis was implemented. The studies served as a test case to explore the potential of large scale correlation studies in the metabolomics field. The work was central to the PhD thesis of Dr. Dirk Steinhauser and was applied in co-supervision with Prof. Dr. Thomas Altmann to the investigation of the subtilase gene family (Rautengarten et al. 2005) and to the identification of brassinosteroid-related genes (Lisso et al. 2005).

BIOINFORMATICS

Vol. 20 no. 12 2004, pages 1928–1939
doi:10.1093/bioinformatics/bth182

Hypothesis-driven approach to predict transcriptional units from gene expression data

Dirk Steinhauser, Björn H. Junker, Alexander Luedemann, Joachim Selbig and Joachim Kopka*

Max Planck Institute of Molecular Plant Physiology, 14476 Golm, Germany

Received on August 25, 2003; revised on February 24, 2004; accepted on February 25, 2004
Advance Access publication March 25, 2004

ABSTRACT

Motivation: A major issue in computational biology is the reconstruction of functional relationships among genes, for example the definition of regulatory or biochemical pathways. One step towards this aim is the elucidation of transcriptional units, which are characterized by co-responding changes in mRNA expression levels. These units of genes will allow the generation of hypotheses about respective functional interrelationships. Thus, the focus of analysis currently moves from well-established functional assignment through comparison of protein and DNA sequences towards analysis of transcriptional co-response. Tools that allow deducing common control of gene expression have the potential to complement and extend routine BLAST comparisons, because gene function may be inferred from common transcriptional control.

Results: We present a co-clustering strategy of genome sequence information and gene expression data, which was applied to identify transcriptional units within diverse compendia of expression profiles. The phenomenon of prokaryotic operons was selected as an ideal test case to generate well-founded hypotheses about transcriptional units. The existence of overlapping and ambiguous operon definitions allowed the investigation of constitutive and conditional expression of transcriptional units in independent gene expression experiments of *Escherichia coli*. Our approach allowed identification of operons with high accuracy. Furthermore, both constitutive mRNA co-response as well as conditional differences became apparent. Thus, we were able to generate insight into the possible biological relevance of gene co-response. We conclude that the suggested strategy will be amenable in general to the identification of transcriptional units beyond the chosen example of *E.coli* operons.

Availability: The analyses of *E.coli* transcript data presented here are available upon request or at http://csbdb.mpimp-golm.mpg.de/

Contact: Steinhauser@mpimp-golm.mpg.de

*To whom correspondence should be addressed.

INTRODUCTION

Public availability of complete genome sequence information (Perna *et al.*, 2001; Blattner *et al.*, 1997) inspired and facilitated the development and utilization of multi-parallel techniques for monitoring the complete cellular inventory. Recent results of these technologies are made available in biological databases that harbour genomic data, gene expression data and information about proteins, metabolites and metabolic pathways. This information will become an empirical basis of understanding the paradigm of life's complexity pyramid (Oltvai and Barabási, 2002). Functional assignment of novel genes, which were discovered by genome sequencing projects, will continue to be the most important goal of the genomic research area (Vukmirovic and Tilghman, 2000). One of the central challenges in computational biology is the discovery of regulatory networks that control gene transcription in biological model systems.

Accumulation of publicly available microarray data led to the development of a range of computational approaches to retrieve biologically meaningful information from co-responding changes of mRNA expression. A variety of computational approaches were previously applied to predict operons from full genome and transcriptome information (Zheng *et al.*, 2002; Moreno-Hagelsieb and Collado-Vides, 2002; Ermolaeva *et al.*, 2001; Yada *et al.*, 1999). Tjaden *et al.* (2002) utilized *Escherichia coli* microarrays to monitor expression of both coding and non-coding intergenic regions. Hidden Markov models were applied to estimate gene boundaries. However, the lack of intergenic probes in routine microarray experiments currently restricts the general application of this approach. Yamanishi *et al.* (2003) applied a generalized kernel canonical correlation analysis to group genes, which share similarities with respect to position within the genome and gene expression. However, this method was restricted to subsets of *E.coli* genes that comprised known metabolic pathways. Bockhorst *et al.* (2003a,b) successfully predicted operons by applying models of transcriptional units to gene sequence and expression data (Bockhorst *et al.*, 2003a) and reported an approach based on Bayesian networks (Bockhorst *et al.*, 2003b). Sabatti *et al.* (2002) re-addressed

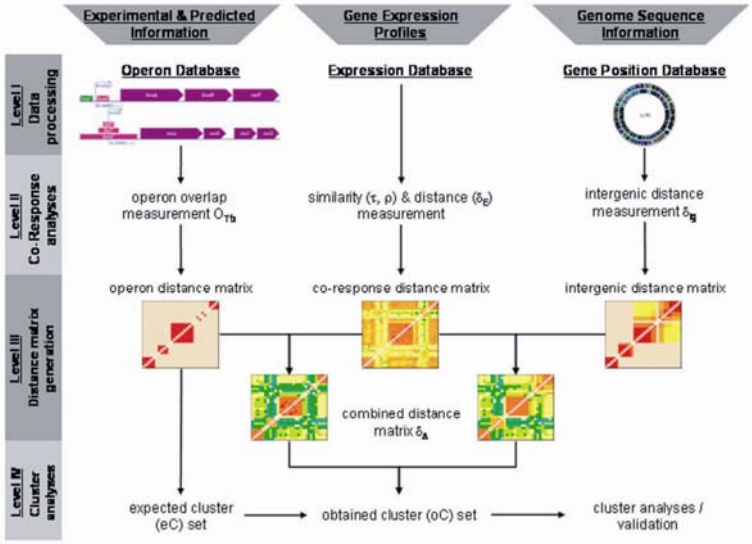


Fig. 1. Flow scheme of adopted data management and processing strategy. Initial data (level I) were converted into pairwise gene distance measures (level II), which were subsequently normalized, arranged into distance matrices and matrix combinations (level III). Finally matrices were subject to hierarchical cluster analyses and the resulting cluster memberships of genes was compared (level IV).

operon prediction by Bayesian classification and described required features.

Till today, no attempt has been made to assign transcriptional units by hierarchical clustering and co-clustering. Here we present a strategy (Fig.1) that was designed to monitor occurrence of constitutive and conditional usage of transcription units in independent gene expression profiling experiments. Co-clustering was demonstrated to be a versatile tool to investigate how prokaryotic genome organization is reflected within compendia of gene expression data. Moreover, we show effects of additional, currently unknown mechanisms on gene co-response, which will be targets of further experimental verification.

METHODS

Data management and processing

Data management, subsequent data integration and processing is summarized within a flow scheme (Fig. 1). Primary data were operon annotations, gene expression data and gene positions within the *E.coli* genome (Fig. 1, level I). These primary data were converted into pairwise gene distances: (1) common operon membership of gene pairs, (2) Pearson's linear correlation coefficient (ρ), Kendall's correlation coefficient (τ), or Euclidean distance (δ_E) of gene pairs as detected within transcript profiles, and (3) pairwise intergenic nucleotide distance of genes (Fig.1, level II). Pairwise gene distances were normalized and combined into distance matrices (Fig.1, level III). Finally, different combined and non-combined distance matrices were subject to hierarchical (co-)clustering and subsequent comparison of cluster memberships of genes. (Fig. 1, level IV). The procedures were as follows:

Data source and pre-processing

Two data sources of *E.coli* gene expression profiles were used (Table 1). The first dataset, called M45 in the following, was derived from the Stanford Microarray Database using ratio 2 values only [SMD, (Sherlock *et al.*, 2001)]. M45 contained 74 expression profiles encompassing 4264 genes, which were analysed by colour-coded cDNA hybridization technology. M45 comprised experiments mainly related to aminoacid metabolism (Khodursky *et al.*, 2000), DNA metabolism (Courcelle *et al.*, 2001), and RNA decay (Bernstein *et al.*, 2002). M45 was quality checked as recommended by the SMD tutorial (http://genome-www5.stanford.edu/help/index.html) and consequently had 43% missing data. In order to reduce the number of missing data two subsets were generated, each of which had only 6% of missing data. M4501 was designed to maximize the number of experiments and comprised 43 experiments, which

Table 1. Overview of the expression of datasets and subsets

Dataset		M45[a]	M4501[a]	M4502[a]	M96[b]	M96A[b]	M96B[b]
Microarray platform[c]		cc	cc	cc	cc, on	on	cc
Number of experiments		74	43	34	66	16	50
Number of genes		4264	929	1845	4241	4345	4290
% of missing data		43	6	6	0	0	0
Experiment categories	Subcategories						
Miscellaneous	Miscellaneous	4	2				
Amino acid metabolism	Tryptophan	29	13	9			
DNA metabolism	UV radiation	15	10	8			
RNA decay	RNA	10	10	9			
	Rifampicin	12	8	8			
Stress/Antibiotics	Acid shock				8	1	7
	Cipro				8	1	7
	Cold shock				4		4
	Heat shock				1	1	
Growth curve	Growth				10	4	6
Media comparison	Media	4			8		8
Strain/mutant analysis	Strains				27	9	18

[a]Stanford Microarray Database (SMD).
[b]ASAP database.
[c]Technology platforms: cc (colour-coded EST hybridization), on (oligonucleotide hybridization).

were described by 929 genes. M4502, designed to maximize the number of genes, comprised 34 experiments, which were described by 1845 genes. Data were normalized and log-transformed as suggested (Sherlock *et al.*, 2001).

The second dataset, M96, was from the ASAP collection (Glasner *et al.*, 2003) and originated from colour-coded cDNA hybridization and oligonucleotide microarray technology (Affymetrix, Santa Clara, CA, USA). M96 consists of 66 experiments comprising 4241 genes, was log-transformed and lacked missing data. The transcript profiling experiments of M96 covered miscellaneous stress treatments, application of antibiotics, comparison of media and growth conditions, as well as characterization of mutants (Allen *et al.*, 2003). Two subsets of M96 were formed to separate profiles from different technology platforms, M96A and M96B (Table 1).

Topological overlap matrices of operon annotations

The operon annotations and predictions were retrieved from Regulon (Salgado *et al.* 2001; http://www.cifn.unam.mx/Computational_Genomics/regulondb/), EcoCyc (Karp *et al.* 2002, http://biocyc.org/ecoli/) and KEGG (http://www.genome.ad.jp./kegg/) databases as well as from Moreno-Hagelsieb and Collado-Vides (2002). All operon assignments were combined into two matrices. First, overlapping and conflicting operon annotations from different sources were maintained within a topological operon overlap matrix, O_{Tu}, according to Ravasz *et al.* (2002). O_{Tu} harbours the combined hypotheses of the maximum possible number and size of operons in *E.coli*. Second, an intersection operon matrix, IS_{tu}, was constructed to comprise the hypotheses of the non-ambiguous and commonly accepted minimum set of *E.coli* operons.

Genome information and weighted intergenic distance (δ_{ig}^{ω})

The EcoCyc database (Karp *et al.*, 2002) was accessed to retrieve gene position and intergenic distances. Intergenic distance of any two genes was defined as follows: genes were separated into the two co-linear groups positioned on the two opposite circular genomic strands of *E.coli*. Within each group, the smaller of the two possible sums of all non-coding nucleotides (nt) in-between any two genes was calculated. The distance of overlapping genes on the same strand was set at zero. A weighted intergenic distance (δ_{ig}^{ω}) was generated from nt distances by normalization to $0 < \delta_{ig}^{\omega} < 1$. Above an nt threshold δ_{ig}^{ω} was set to 1, below this threshold δ_{ig}^{ω} was calculated by dividing non-coding intergenic nt by the respective threshold value. Four threshold values were chosen, 2×2250 nt, 2×7250 nt, $2 \times 70\,000$ nt and $2 \times 655\,596$ nt. The rationale for choosing these thresholds is reported below.

Co-response matrices

Pearson's product moment linear correlation (ρ), non-parametric Kendall's coefficient of rank correlation (τ) without correction for ties (Sokal and Rohlf, 1995) and Euclidean distance coefficients (δ_E) (Mirkin, 1996) were applied to log-transformed gene expression data. The significance of correlation was tested as recommended by Sokal and Rohlf (1995). In order to generate normalized distance

matrices, correlation coefficients were converted according to Mirkin (1996) and Sokal and Rohlf (1995). The largest distance was assigned to negative Pearson's or Kendall's correlation coefficients. The converted distances were marked with the index of used correlation coefficient (e.g. δ_τ). Then all distance measures were normalized to the maximum distance value. Thus, all resulting normalized distances (δ^ν) were in the range of $0 \leq \delta^\nu \leq 1$.

Joining function (λ_ψ) and combined distance (δ_Δ)

Normalized distance matrices were combined as suggested by Hanisch *et al.* (2002) applying a modified function, (λ_ψ), Equation (1), extended to n dimensions. The resulting combined distance function δ_Δ of each gene pair, g_i, g_j, with $\psi \in \{\delta^\nu_\rho, \delta^\nu_\tau, \delta^\nu_E, \delta^\nu_{OTtu}, \delta^\omega_{ig}, \ldots\}$, was defined as follows:

$$\delta_\Delta(g_1, g_2) = \frac{1}{n}\sum_{\psi=1}^{n}\left[\lambda_\psi\left(g_i, g_j\right)\right]$$

$$\text{where } \lambda_\psi(g_i, g_j) = \frac{1}{1 + e^{-\{[\delta_\psi(g_i,g_j) - \nu_\psi]/s_\psi\}}} \qquad (1)$$

For co-response distances the control parameters of the shape of the logistic curve (ν_ψ, S_ψ) were adjusted to the median of distance distribution (ν_ψ) and to a moderate slope of $S_\psi = \nu_\psi/6$. The control parameters of the δ^ω_{ig}S were adjusted to $\nu_\psi = \nu_{\psi\text{weigting}} - \nu_{\psi\text{correction}} = 0.5 - 0.17578 = 0.32422$ and to $S_\psi = (\nu_{\psi\text{weighting}}/6) = 0.08$. The correction term $\nu_{\psi\text{correction}}$ was empirically determined by fitting $\nu_{\psi\text{weighting}}$ to the formula described above and by the setting of parameters to $\lambda_{ig} = 0.9$ and $\delta^\omega_{ig} = 0.5$. For the generation of O_{Ttu} the parameters were set to $\nu_\psi = 0.5$ and $S_\psi = \nu_\psi/6$. An a priori weighting of 50% ($n = 2$) was assigned to combine O_{Ttu} or δ^ω_{ig} with distance matrices describing transcriptional co-response. Two transcriptional co-response matrices of each expression dataset were combined, i.e. 25% weight was given to either a normalized distance matrix based on Pearson's or Kendall's coefficients and residual 25% weight was assigned to the normalized Euclidian distance matrix (see above).

Hierarchical cluster analysis and cluster validation

For the classification of genes, the unweighted average linkage-clustering algorithm (UPGMA) was applied to normalized distance matrices (Mirkin, 1996). Expected operon clusters (eC) were generated by the use of δ_{OTtu} of the operon overlap matrix (see above). Cluster validation was performed by measuring the degree of correspondence between the expected cluster (eC) and the obtained cluster (oC). In detail, the cluster-specific match coefficient (CMC) reflects the ratio of elements, i.e. genes, from eC that are observed to occur within oC. For example, if eC=oC, obtained clusters perfectly match expectations. The combined match coefficient (MC) was defined at selected clustering heights as the sum of all CMCs divided by the number of expected clusters. MC represents the portion of all genes that were found to belong to expected operons, for example, if MC = 1.0, all genes were found to group into respective expected clusters. The cluster-specific reassignment coefficient (CRC) is the ratio of those genes that are not expected to occur in eC as compared to the genes that are correctly grouped into oC. CRC is indicative of the portion of novel genes that were unexpectedly assigned to any given cluster. The sum of all CRCs at a specific clustering height divided by the number of expected clusters yields the reassignment coefficient (RC). For example, if RC = 1.0, the number of mis-assigned genes is equal to the number of correctly assigned genes.

Statistical analysis and software

The Mantel test and respective analysis of variance, the non-parametric Kruskal–Wallis and two-sample Wilcoxon rank sum tests, tests of homogeneity as well as parametric three-way factorial ANOVA were computed as described by Sokal and Rohlf (1995). The Cramer test was performed according to Baringhaus and Franz (2004). All statistical tests were applied to iterated random selections of data subsets.

Computations were performed using the statistical software environment R (http://www.r-project.org) version 1.6.1. and 1.6.2 with the libraries 'mva', 'exactRankTests', 'vegan', 'e1071', 'tseries', 'ctest' and 'cramer'. Calculations were executed with PERL scripts.

RESULTS

The goal of this work was to investigate how the prokaryotic genome organization, namely polycistronic operon structure, is reflected within different compendia of gene expression profiles from *E.coli* and whether functional linkage of genes can be detected by clustering technologies. We selected the prokaryotic operon structure because genes that are co-regulated in physical units of common polycistronic messenger RNA (mRNA) can be expected to have a high correlation in transcriptome analyses independent of the nature of underlying biological experiments. This strong co-regulation should allow precise classification by clustering technologies irrespective of the nature of distance measure applied. However, initial attempts to retrieve clusters of genes that constitute operons within combined sets or subsets of M45 and M96 failed.

Operon classification by clustering

The co-response matrices of four data subsets M4501, M4502, M96A and M96B were generated using Kendall's τ, Pearson's ρ and Euclidean distance δ_E. These matrices were subject to HCA and subsequently cluster membership of genes was compared to expected clusters as represented by O_{Ttu}. The criterion for clustering quality was MC (cluster match coefficient), and the criterion for gene mis-assignment was RC (reassignment coefficient; refer to the earlier section 'hierarchical cluster analysis and cluster validation'). Only results

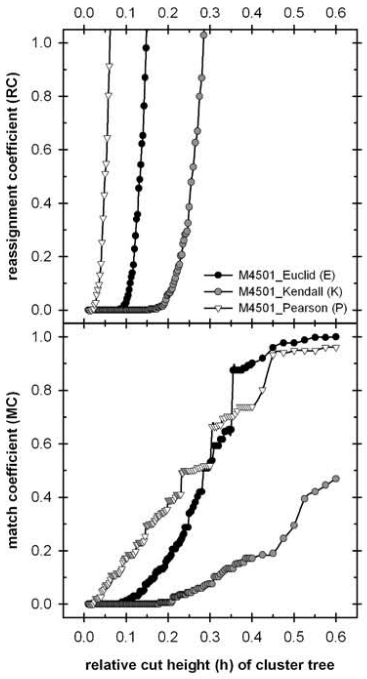


Fig. 2. Plot of match coefficient (MC) (bottom) and the reassignment coefficient (RC) (top) resulting from HCA of distance matrices from dataset M4501 at increasing relative heights (h). RC is shown in the range of 0 to 1.0 with 1.0 representing 50% novel gene assignments. Inset shows applied distance measures.

of M4501, which were representative of all other data subsets, are shown (Fig. 2). Irrespective of the applied distance measure, only a minority of genes was correctly assigned to expected operons, e.g. MC < 0.08, when accepting 50% mis-assignments (RC = 1.0) (Fig. 2).

The analysis of best pairwise gene associations according to Kendall's τ, Pearson's ρ (data not shown) or Euclidean distance δ_E demonstrated that only 50% of those gene associations, which were expected to be the result of genes belonging to the same operons, ranked among the top 5–10% (Fig. 3). We tested whether this observation was caused by applying O_{Tu}, which included ambiguous as well as predicted operon definitions. For this purpose we applied IS_{tu}, i.e. we restricted our analysis to the minimum intersection of available operon annotations (refer to the section 'topological overlap matrices of operon annotations'). MC however, only increased approximately 2-fold at RC = 1.0 and frequency distributions of relative rank positions were independent of choice of either O_{Tu} or IS_{tu} as was supported by a non-parametric Cramer test. Therefore, we had to assume factors other than precision of current operon annotation, which either cause absence of expected pairwise gene associations or which are caused by other regulatory mechanisms of coordinated gene expression, such as transcription factors or mRNA processing. We ruled out an artefact caused by the choice of transcript profiling technology because datasets M96A, and M96B did not show fundamental differences (Fig. 3). In the following we first characterize the nature of the datasets, unravel properties, which obscure operon units and demonstrate that co-clustering and the use of data subsets allowed overcoming this inherent problem of transcriptome analyses.

Properties of transcript datasets

We performed a comparison of the datasets M45 and M96 by applying hierarchical cluster analysis to a δ_E association matrix of the respective compendium experiments (Fig. 4). In dataset M4501 the majority of nodes are in the heterogeneity range of $0 \leq h_e \leq 0.2$ and experiment grouping strongly reflects the nature of underlying biological experiments (Fig. 4a). Similar results were obtained from either M45 or subset M4502 (data not shown). M96 exhibited higher inherent heterogeneity, $0.2 \leq h_e \leq 0.4$. Biological experiments were partially reflected by clustering, but experiments from different technology platforms were clearly separated (Fig. 4b). Non-parametric analysis of variance by Mantel testing revealed a highly significant ($\tau = 0.5668$, $P \ll 0.001$) difference of variance between the experiments of different technology platforms within M96, as measured by median Kendall's τ association; in detail, among experiments of Affymetrix technology $\tau_{median} = 0.558$, among experiments of cDNA technology $\tau_{median} = 0.453$ and in between experiments of different technology platforms $\tau_{median} = 0.222$. Therefore, dividing M96 into M96A and M96B according to technology platform was justified.

Analysis of gene pair association within the different datasets revealed significantly different data structures. The Kendall's τ distribution of gene pair association in M4501 ($\tau_{median} = 0.452$) and M4502 ($\tau_{median} = 0.537$) exhibited strong shifts to positive values, whereas M96A ($\tau_{median} = 0.075$) and M96B ($\tau_{median} = 0.068$) were centred approximately to zero (Fig. 5a). In all datasets, we observed more significant positive gene associations than significant negative associations. Furthermore, even though the datasets appeared to be of a highly diverse structure, the datasets tested positive for the presence of common gene pair associations, when the Mantel test was applied to compare the gene co-response matrices. In addition, test of homogeneity applied to gene pair associations from the above matrices revealed homogeneity levels of 47.8–90.1%. Thus, all datasets contained portions of similar information on pairwise gene associations. This observation was an incentive to pursue subsequent comparative analyses.

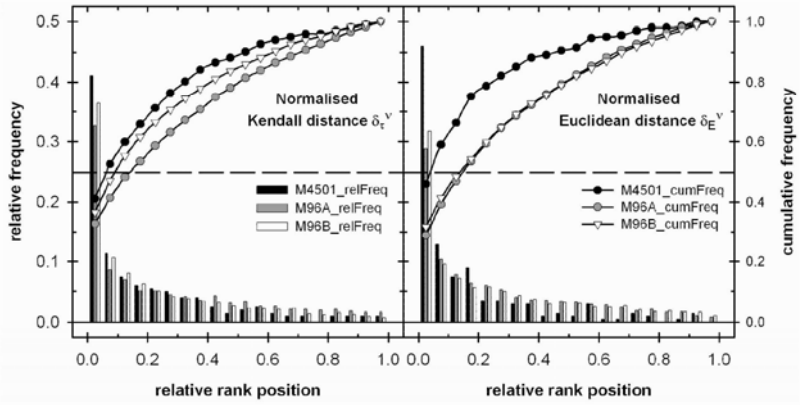


Fig. 3. Histogram plots and cumulative frequency of the relative rank distributions of all gene pairs, which belong to same operons. Best ranking was according to normalized Kendall distance δ_τ^v and normalized Euclidean distance δ_E^v. Dashed lines mark 50% cumulative frequency. Relative rank was rank of gene divided by total number of genes available in respective dataset, M4501, M96A and M96B.

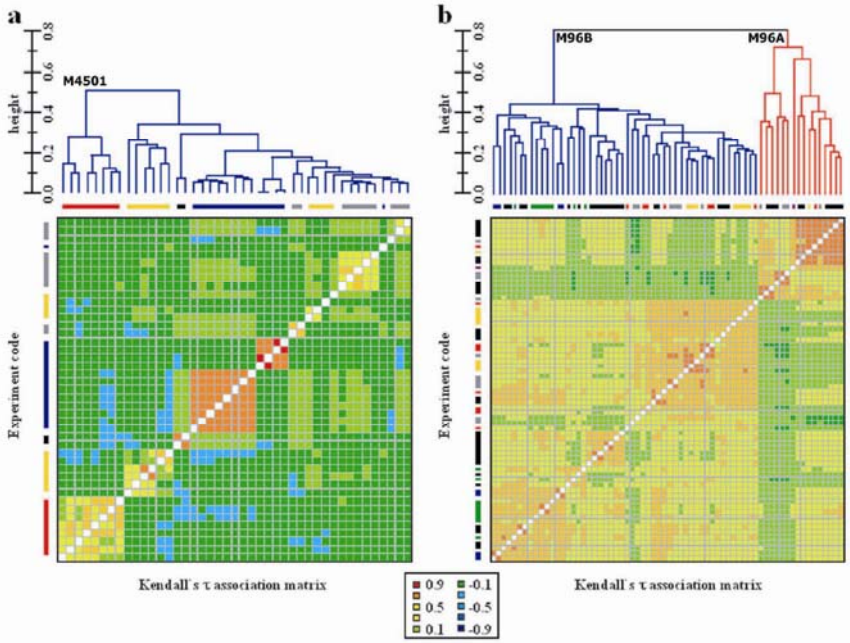


Fig. 4. Comparison of dataset M4501 and M96 by hierarchical clustering of experiments applying Euclidean distance δ_E (top) to a Kendall's τ association matrix (bottom). The experiment categories are colour-coded to the left: M4501 (4a), RNA decay—Rifampicin (red), RNA decay—RNA (yellow), miscellaneous (black), amino acid metabolism—tryptophan (blue), and DNA metabolism—UV radiation (grey); M96 (4b) cold-shock (blue), various strain/mutant characterization (black), media comparison (green), acid-shock (yellow), antibiotic (red), heat-shock (purple), and growth curve (grey).

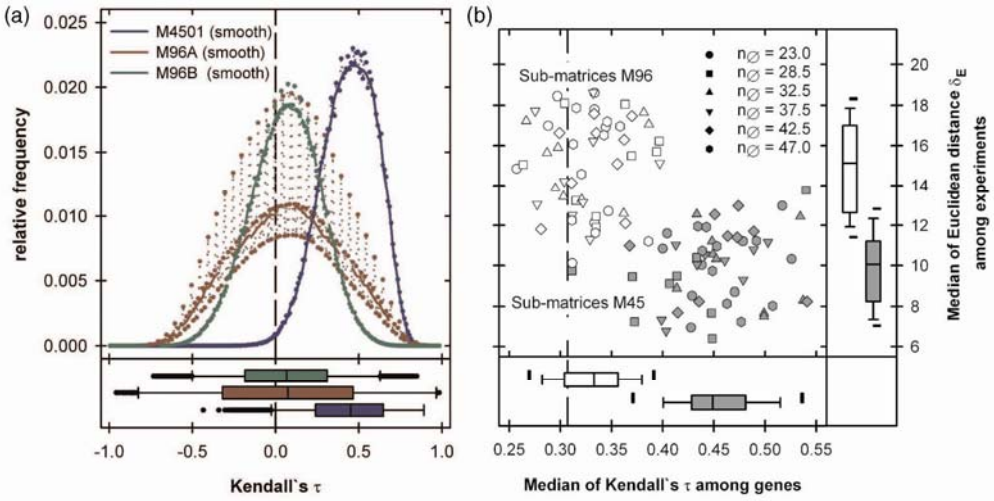


Fig. 5. (**a**) Histograms (top) and box-plots (bottom) of all gene pair association from three Kendall's τ correlation matrices, M4501 (blue), M96A (brown) and M96B (green). The bold lines represent smoothed frequency distributions. (**b**) Scatter-plots and box-plots of median Euclidean distances δ_E representing heterogeneity among experiments and median Kendall's τ of pairwise gene associations from subsets of gene expression data, marked according to source, M45 (grey) and M96 (white). Subsets were created at random, comprising approximately 200 genes each, and were characterized by average numbers of gene pairs (n_ϕ, inset). The dashed vertical line represents the critical significance threshold of gene associations for $n = 22$ at probability $\alpha = 0.05$.

Distributions of gene associations

A major concern of our investigation was the positive shift of the distribution of Kendall's τ gene pair associations evident within M4501 (Fig. 5a) and M4502 (data not shown). We investigated whether the choice of gene expression experiments and experimental diversity (Fig. 4) had an impact on the shift of Kendall's τ distribution of gene pair association (Fig. 5a). In other terms we asked whether prevailing positive associations in dataset M4501 might reflect higher suitability of this dataset to investigate prokaryotic operon organization as compared to M96.

For this purpose, we created 10 random gene sets, comprising approximately 200 genes each. Genes were chosen only once and had to be present in both datasets. Transcript data of each gene set were extracted separately from M96 and M45. Experiments of each dataset were chosen at random to comprise data subsets that had average numbers of gene pairs as follows: 47.0, 42.5, 37.5, 32.5, 28.5 and 23.0. The numbers of gene pairs were smaller than the numbers of experiments, because of missing data in M45. Heterogeneity among experiments of each selection was determined by median δ_E (Fig. 5b). Our analyses revealed a relation of median δ_E and median Kendall's $\tau(\tau_{median})$, e.g. reduced heterogeneity of experiments coincided with a positive shift of τ_{median} of gene pair associations. The portion of significant positive gene pair associations was increased in M45 (Fig. 5b) as compared to M96. Subsequently, by application of parametric ANOVA on τ_{median} distributions we tested the factor that might influence the above observation, namely choice of gene subsets, number of gene pairs, i.e. 23.0–47.0 or data source, i.e. M96 or M45. No first- and second-order interactions among these factors were found. Only the data source had a significant influence, ANOVA: $F_s > 293$, $P = 2.50 \times 10^{-15}$. A subsequent non-parametric Kruskal–Wallis test ($P < 7.00 \times 10^{-06}$) substantiated this finding. In conclusion, the shift of τ_{median} distributions of pairwise gene associations was inherent to datasets and not biased by different numbers of experiments or choice of technology (see above).

However, on comparing Kendall's τ distribution of all genes from M96, either M96A or M96B (Fig. 5a), with the distributions of gene subsets from M96 (Fig. 5b), we observed that the τ_{median} shifted from 0.053 (M96, 4241 genes) to approximately 0.25–0.4 (M96, approximately 200 genes each). Thus, shifts of τ_{median} can be caused by the choice of gene subsets.

In order to investigate this observation, we subdivided gene associations into one group that represents gene associations by operon structures (type I associations) and a second much larger group of all gene associations that do not describe operon structures (type II associations). Comparative analysis of Kendall's τ distributions indicated fundamentally different

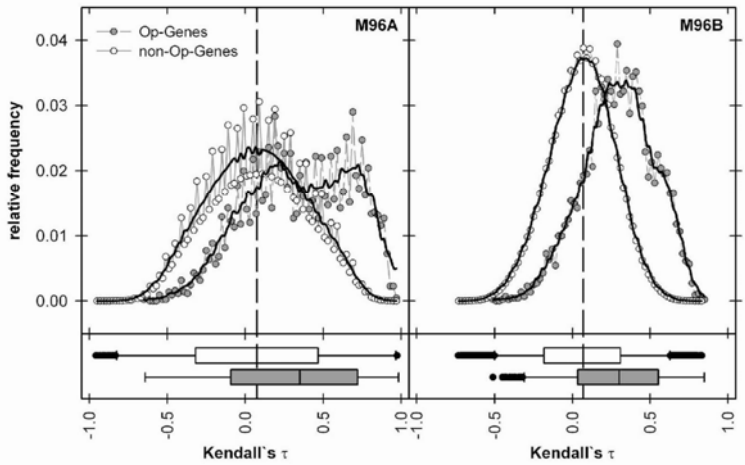


Fig. 6. Kendall's τ distribution of type I associations, i.e. related to operon structure, and type II associations, i.e. not related to operon structure apparent in data subsets M96A and M96B. Dashed lines indicate median of Kendall's τ for the combined type I and type II associations. The bold lines represent smoothed frequency distributions.

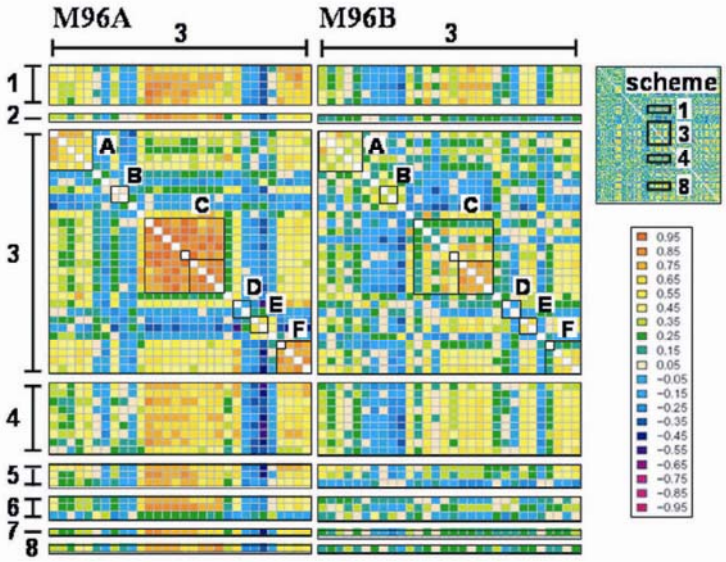


Fig. 7. Colour-coded partial visualization of Kendall's τ association matrix of genes from the chromosomal section nt 15.995 – 16.718. Nearest neighbour associations (section 3) and selected distant associations (sections 1, 2, 4–8) from data subsets M96A (left) and M96B (right) are shown. Sections were: (1) operon *cyoABCDE*, (2) nmpC, (3) region *ybgI* to *tolA*, (4) region *flgB* to *flgJ*, (5) *trpD*, *trpE* and *trpL*, (6) operon *ivbL-ilvBN*, (7) *metH* and (8) *malE*. Section 1 represents operons: A., *ybgIJKL-nei*; B, *ybgPQ*; C, *sdhCDAB-b0725-sucABCD*; D, *hrsA-ybgG*; E, *cydAB*; and F, *yvgC-tolQRA*. Alternative or ambiguous operon annotations are marked by boxes.

properties (Fig. 6). While type II associations were approximately normal distributed and exhibited almost equal numbers of high positive and negative Kendall's τ, type I associations were predominantly positive as was supported by Wilcox rank sum testing and shown previously with an independent dataset (Sabatti *et al.*, 2002). A portion of 75% of type I associations were overlapping with significant and positive type II associations and thus, the high numbers of type II associations were obscured (Fig. 7). In addition to previous observations we found evidence of bi-modal type I associations (Fig. 6; M96A), indicative of a mechanism that apparently uncouples associations based on polycistronic mRNA. We define these associations as type III, i.e. those associations that according to operon annotations were expected to be significantly positive, but were found to be non-significant or had even negative and significant Kendall's τ. The type III associations were variable in numbers, as was exemplified by M96A and M96B (Fig. 6), highly operon specific, and dependent on the choice of experiments (Fig. 7).

Operon classification by co-clustering

We demonstrated earlier that type I and type II associations cannot be differentiated without utilizing additional knowledge. Co-clustering was suggested to integrate multiple information sources for cluster analyses (Hanisch *et al.*, 2002). We modified this technology to allow overlay of operon annotation, intergenic distance and transcriptional co-response data into combined matrices and subsequent HCA (Fig. 1). We first adjusted co-clustering to enforce classification of genes belonging to the same operons (Fig. 8; OpOVLP). Choice of joining function and weighting was as described earlier (refer to the section 'joining function and combined distance'). Consequently MC approximated 1.0 and RC was 0.0 at clustering height 0.5. The stringency of co-clustering was set to merge even those genes that exhibited negative correlation to other members of these operons into correct operon clusters. Representative results from M4501 are shown (Fig. 8). Instead of adjusting joining function and weighting of co-clustering to allow novel association to operon clusters or rule out previous annotations, we maintained settings but substituted O_{Ttu} for matrices, which described physical proximity and co-linearity of genes on chromosomal strands. The maximum nt allowed applied to construct matrices of weighted intergenic distances (δ_{ig}^{ω}) was optimized to approximate the co-clustering results obtained by O_{Ttu} (Fig. 8). Use of the maximum nt distance threshold, namely 655 596 nt representing the average of non-coding nt divided by 2 of both DNA strands, improved results markedly, MC > 0.3 at RC 1.0 as compared to direct clustering (Fig. 2). Applying a 70 000-nt threshold doubled MC at RC = 1.0. The number of 70 000 nt is equal to the maximum of non-coding nt observed in between any set of 16 co-linear and adjacent genes. The choice of 16 genes was motivated by the largest known operon of *E.coli*, which has 15 genes. The choice of further thresholds, 2250 and 7250 nt, was equal to the maximum length of 95% or of all *E.coli* genes, respectively. Application of these thresholds mimics MC traces obtained with O_{Ttu} (Fig. 8), whereas RC, i.e. the percentage of novel reassignments, can be fine-tuned by selecting the threshold number of nt (data not shown). MC traces obtained with equal settings were mostly independent of datasets. For example, when we applied the 2250 nt threshold, MC only ranged from 0.87 to 1.0 at RC = 1.0. In contrast, RC traces varied widely. RC traces of M96A and M96B had steeper slopes as compared to M4501 and similarly M96B had a steeper RC slope as compared to M96A (data not shown). This observation was indicative of RC slopes increasing with heterogeneity of experiments, $h_{M4501} < h_{M96A} < h_{M96B}$.

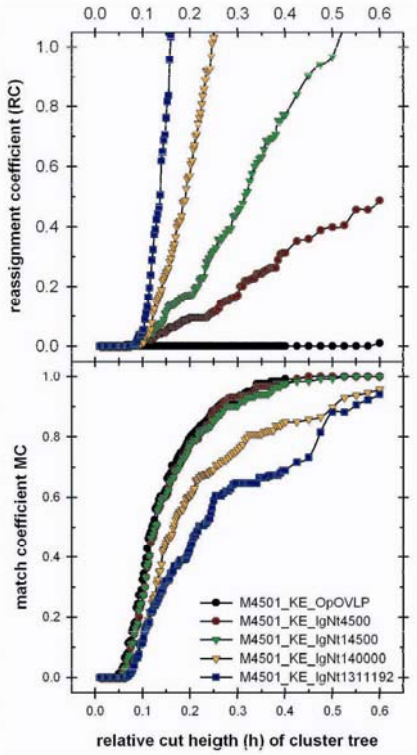


Fig. 8. Plot of match coefficient (MC) (bottom) and the reassignment coefficient (RC) (top) resulting from HCA of distance matrices from dataset M4501 at increasing clustering heights (h). RC is shown in the range of 0.0–1.0 with 1.0 representing 50% mis-assigned genes. Kendall's τ distance measure was applied. Inset shows different co-clustering: OpOVLP forced co-clustering by O_{Ttu}.

Table 2. Kendall's τ association of operon *sdhCDAB-(b0725)-sucABCD* and co-clustering using combined matrices (δ_Δ) of δ_{ig}^{ω} at 2250 nt threshold with Kendall's τ matrices derived from M96A and M96B

	b0721	b0722	b0723	b0724	b0725	b0726	b0727	b0728	b0729	Type	RIC[a]	rel h[b]	RC[c]
M96A													
b0721			avg.	0.84				operon avg.			2	0.00	0.07
b0722	0.88							0.80			2	0.00	0.07
b0723	0.83	0.92									4	0.03	0.09
b0724	0.75	0.80	0.88								3	0.02	0.08
b0725	0.88	0.90	0.88	0.80						I	1	0.00	0.06
b0726	0.80	0.85	0.83	0.78	0.85			avg.	0.81		1	0.00	0.06
b0727	0.63	0.68	0.77	0.85	0.72	0.73					4	0.03	0.09
b0728	0.72	0.77	0.85	0.83	0.77	0.82	0.92				5	0.15	0.13
b0729	0.72	0.73	0.82	0.87	0.73	0.78	0.78	0.80			4	0.03	0.09
M96B													
b0721			avg.	0.81				operon avg.			5	0.69	0.26
b0722	0.21							0.41			3	0.05	0.10
b0723	0.17	0.09									5	0.69	0.26
b0724	0.16	0.72	0.13								3	0.05	0.10
b0725	0.21	0.47	−0.02	0.45						I	4	0.17	0.13
b0726	0.11	0.64	0.09	0.66	0.43			avg.	0.73		2	0.01	0.08
b0727	0.06	0.57	0.19	0.65	0.34	0.74					1	0.00	0.07
b0728	0.14	0.62	0.03	0.63	0.38	0.82	0.72				2	0.01	0.08
b0729	0.13	0.61	0.19	0.71	0.41	0.70	0.74	0.68			1	0.00	0.07

[a]Rank of merger into operon cluster.
[b]Relative height at merger.
[c]Reassignment coefficient of dataset at merging height.

Analysis of operon structures

The clustering results can now be used to investigate operon annotations as well as to compare and validate transcriptional co-response in different datasets, e.g. under different experimental conditions. In the following, we analyse two exemplary *E.coli* operons in detail using combined matrices (δ_Δ) of δ_{ig}^{ω} at 2250 nt threshold with Kendall's τ matrices of the different transcript datasets. Gene clusters were created with clustering heights set at RC = 0.5, i.e. allowing a portion of 33.3% novel gene assignments. Because M96 had a more complete representation of the full genome and most of the genes discussed below were only partially represented in M45 we focussed our investigations on M96.

The complex operon *sdhCDAB-(b0725)-sucABCD* (Fig. 7, section 3C) codes for succinate dehydrogenase (*sdhCDAB*), components of the 2-oxoglutarate dehydrogenase complex (*sucAB*) and parts of succinyl-CoA synthetase (*sucCD*) (Cunningham and Guest, 1998). This operon allows control of a major constituent of the tricarboxylic acid cycle in central metabolism. Prevoius experimental characterization identified two promotor elements, P_{sdh} and the internal P_{suc}, as well as the co-transcription of the entire operon into a 'nine-cistron' mRNA (Cunningham and Guest, 1998). RegulonDB assigned three possible transcripts: (1) the entire *sdhCDAB-b0725-sucABCD* mRNA, (2) a *sucABCD* mRNA and (3) a single-gene *b*0725 transcript. In contrast, EcoCyc annotates only the first two mRNAs. Co-clustering of δ_Δ of M96A supports activity of P_{sdh} and expression of the full 'nine-cistron' transcript (b0721–b0729). All genes of this operon were merged into one cluster (Table 2) and respective type I gene associations were all significant ($P \ll 0.001$). Dataset M96B and underlying experiments revealed differential regulation of transcriptional co-response (Fig. 7, Table 2). This dataset supports transcriptional activity of P_{suc}, as is evident through significant type I associations of the *sucABCD* operon (b0726–b0729). All other expected associations were type III, e.g. absent or strongly reduced. Indeed, we can propose one possible underlying mechanism of type III associations: differential use of stacked promoters in *E.coli*. These promoters control overlapping subsets of genes, which can be differentially controlled under varying experimental conditions, as was demonstrated.

As a second example, we selected operon *nlpB-dapA*, which according to RegulonDB may be controlled by P_{dapA} (Salgado *et al.*, 2001). *NlpB* (b2477) encodes lipoprotein-34 and *dapA* (b2477) code for dihydrodipicolinate synthase (EC 4.2.1.52). Co-cluster analyses of M96A and M96B merged both genes into one cluster (Table 3). In addition, adjacent *purC* (b2476), which encodes a subunit of phosphoribosylaminoimidazole–succinocarboxamide synthase, was assigned to this cluster irrespective of the choice of dataset (Table 3). All associations of *purC* with *nlpB-dapA* were significant. Based on the proximity of *purC* and *nlpB-dapA* we can propose a transcriptional unit *purC-nlpB-dapA*. Two mechanisms

Table 3. Kendall's τ association of transcriptional unit *purC-nlpB-dapA* (operon *nlpB-dapA*) and co-clustering of combined matrices (δ_Δ) of δ_{tg}^{ω} at 2250 nt threshold with Kendall's τ matrices derived from M96A and M96B

	b2476	b2477	b2478	Type	RIC[a]	rel h[b]	RC[c]
M96A							
b2476		avg.	0.81		2	0.06	0.10
b2477	0.88			I	1	0.02	0.08
b2478	0.83	0.72			1	0.02	0.08
M96B							
b2476		avg.	0.59		1	0.02	0.08
b2477	0.52			I	2	0.16	0.13
b2478	0.66	0.59			1	0.02	0.08

[a]Rank of merger into operon cluster.
[b]Relative height at merger.
[c]Reassignment coefficient of dataset at merging height.

may be the cause of this unit, either operon structure or a strong transcriptional control of two adjacent genes by a common transcription factor. The mechanism remains to be investigated by experimental analyses of transcript length.

DISCUSSION

In this paper we chose a hypothesis-driven co-clustering approach for the identification of transcriptional units. As ideal test cases of co-transcribed genes we used operon structures that result in physically linked co-transcription. Furthermore, we applied our approach to three independent sets of biological experiments using an overlap matrix O_{Tu}, which represents the combination of all available annotations of polycistronic *E.coli* operons. We clearly demonstrate the failure of the assumption that polycistronic mRNA inevitably results in high gene-to-gene correlation of transcript measurements. We unravel two major mechanisms that contribute to obscure operon structures within transcript profiles. First, presence of type II associations, which include synergistic (positive Kendall's τ) and antagonistic (negative Kendall's τ) control of distant genes by common transcription factors. These associations dominate in numbers any correlation matrix and overlap with type I associations. Second, type III associations exist. Expected high transcriptional co-response due to operon structures may indeed be conditional, because of known or still undiscovered stacked promoters. An example of the differential use of stacked promoters under different experimental conditions is shown. Moreover, additional mechanisms contributing to type III associations might be functional, such as post-transcriptional mRNA processing and degradation.

We conclude that only recruiting additional information will allow extraction of operon structures from gene expression data. We successfully applied co-clustering technology to include gene distance information and demonstrated that gene distance as suggested earlier by Sabatti *et al.* (2002) can effectively substitute information about known operon annotations (Fig. 8). Furthermore, we show evidence that comparative analyses on data subsets, which describe defined experimental interventions, will be highly informative as compared to global analyses of compendium datasets. The presence of binding sites for multiple transcription factors within many promoter regions as well as the occurrence of stacked promoters driving different gene subsets of operons indicate that many overlapping transcription units may exist and can be used in response to varying stimuli. Our analyses demonstrate differential as well as constitutive use of exemplary transcriptional units. Transcription units were shown to be highly dependent on experimental conditions (Fig. 7, Table 2). We envision that analyses of constitutive activity or conditional use of operons and transcriptional units controlled by transcription factors will be the imminent task of transcriptome analyses and stimulate will lead to further experimental investigations.

ACKNOWLEDGEMENTS

We thank the staff of the SMD and the ASAP database for the establishment of public accessible sources for microarray data as well as all scientists who submitted transcript profile data to these databases and thereby enabled comparative investigations. Furthermore, we thank the Free Software Foundation (FSF) for access to software under the terms of the GNU general public license. We acknowledge L. Krall, A. Fernie and J. Kehr for their critical reading of this manuscript. Furthermore, the comments from the two anonymous referees are gratefully acknowledged.

REFERENCES

Allen,T.E., Herrgård,M.J., Liu,M., Qiu,Y., Glasner,J.D., Blattner,F.R. and Palsson,B.Ø. (2003) Genome-scale analysis of the uses of the *Escherichia coli* genome: a model-driven analysis of heterogeneous datasets. *J. Bacteriol.*, **185**, 6392–6399.

Baringhaus,L. and Franz,C. (2004) On a new multivariate two-sample test. *J. Multivar. Anal.*, **88**, 190–206.

Bernstein,A.J., Khodursky,A.B., Lin.,P.-H., Lin-Chao,S. and Cohen,S.N. (2002) Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarray. *Proc. Natl Acad. Sci., USA*, **99**, 9697–9702.

Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis.N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau, B. and Shao,Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.

Bockhorst,J., Qiu,Y., Glasner,J., Liu,M., Blattner,F.R. and Craven,M. (2003a). Predicting bacterial transcription units using sequence and expression data. *Bioinformatics*, **19**, i34–i43.

Bockhorst,J., Craven,M., Page,D., Shavlik,J. and Glasner,J. (2003b) A Bayesian network approach to operon prediction. *Bioinformatics*, **19**, 1227–1235.

Courcelle,J., Khodursky,A., Peter,B., Brown,P.O. and Hanawalt,P.C. (2001) Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient *Escherichia coli*. *Genetics*, **158**, 41–64.

Cunningham,L. and Guest,J.R. (1998) Transcription and transcript processing in the *sdhCDAB-sucABCD* operon of *Escherichia coli*. *Microbiology*, **144**, 2113–2123.

Ermolaeva,M.D., White,O. and Salzberg,S.L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res.*, **29**, 1216–1221.

Glasner,J.D., Liss,P., Plunkett,G.,III, Darling,A., Prasad,T., Rusch,M., Byrnes,A., Gilson,M., Biehl,B., Blattner,F.R. and Perna,N.T. (2003) ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res.*, **31**, 147–151.

Hanisch,D., Zien,A., Zimmer,R. and Lengauer,T. (2002). Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18**, 145–154.

Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Collado-Vides,J., Paley,S.M., Pellegrini-Toole,A., Bonavides,C. and Gama-Castro,S. (2002) The EcoCyc Database. *Nucleic Acids Res.*, **30**, 56–58.

Khodursky,A.B., Peter,B.J., Cozzarelli,N.R., Botstein,D., Brown,P.O. and Yanofsky,C. (2000) DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in *Escherichia coli*. *Proc. Natl Acad. Sci., USA*, **97**, 12170–12175.

Mirkin,B. (1996) Mathematical Classification and Clustering. (Book Series: Nonconvex Optimisation and Its Application: Volume 11). Kluwer Academic Publishers, London.

Moreno-Hagelsieb,G. and Collado-Vides,J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18**, S329-S336.

Oltvai,Z.N. and Barabási,A.-L. (2002) Life's complexity pyramide. *Science*, **298**, 763–764.

Perna,N.T., Plunkett,G.,III, Burland,V., Mau,B., Glasner,J.D., Rose,D.J, Mayhew,G.F., Evans,P.S., Gregor,J., Kirkpatrick,H.A. *et al.* (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* 0157:H7. *Nature*, **409**, 529–533.

Ravasz,E., Somera,A.L., Mongru,D.A., Oltvai,Z.N. and Barabási,A.-L. (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Sabatti,C., Rohlin,L., Oh,M.-K. and Liao,C. (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res.*, **30**, 2886–2893.

Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millán-Zárate,D., Diaz-Peredo,E., Sánchez-Solano,F., Pérez-Rueda,E., Bonavides-Martinez,C. and Collado-Vides,J. (2001) RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.*, **29**, 72–74.

Sherlock,G., Hernandez-Boussard,T., Kasarskis,A., Binkley,G., Matese,J.C., Dwight,S.S., Kaloper,M., Wenig,S., Jin,H., Ball,C.A. *et al.* (2001) The stanford microarray database. *Nucleic Acids Res.*, **29**, 152–155.

Sokal,R.R. and Rohlf,F.J. (1995) Biometry: The Principles and Practice of Statistics in Biological Research, 3rd edn. W.H. Freeman and Company, New York.

Tjaden,B., Haynor,D.R., Stolyar,S., Rosenow,C. and Kolker,E. (2002) Identifying operons and untranslated regions of transcripts using *Escherichia coli* RNA expression analysis. *Bioinformatics*, **18**, 337–344.

Vukmirovic,O.G. and Tilghman,S.M. (2000) Exploring genome space. *Nature*, **405**, 820–822.

Yada,T., Nakao,M., Totoki,Y. and Nakai,K. (1999) Modeling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models. *Bioinformatics*, **15**, 987–993.

Yamanishi,Y., Vert,J.-P., Nakaya,A. and Kanehisa,M. (2003). Extraction of correlated gene clusters from multiple genomic data by generalised kernel canonical correlation analysis. *Bioinformatics*, **19**, i323–i330.

Zheng,Y., Szustakowski,J., Fortnow,L., Roberts,R. and Kasif,S. (2002) Computational identification of operons in microbial genomes. *Genome Res.*, **12**, 1221–1230.

BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 18 2004, pages 3647–3651
doi:10.1093/bioinformatics/bth398

CSB.DB: a comprehensive systems-biology database

Dirk Steinhauser*,†, Björn Usadel†, Alexander Luedemann, Oliver Thimm and Joachim Kopka

Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm, Germany

Received on June 5, 2004; accepted on June 30, 2004
Advance Access publication July 9, 2004

ABSTRACT

Summary: The open access comprehensive systems-biology database (CSB.DB) presents the results of bio-statistical analyses on gene expression data in association with additional biochemical and physiological knowledge. The main aim of this database platform is to provide tools that support insight into life's complexity pyramid with a special focus on the integration of data from transcript and metabolite profiling experiments. The central part of CSB.DB, which we describe in this applications note, is a set of co-response databases that currently focus on the three key model organisms, *Escherichia coli*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. CSB.DB gives easy access to the results of large-scale co-response analyses, which are currently based exclusively on the publicly available compendia of transcript profiles. By scanning for the best co-responses among changing transcript levels, CSB.DB allows to infer hypotheses on the functional interaction of genes. These hypotheses are novel and not accessible through analysis of sequence homology. The database enables the search for pairs of genes and larger units of genes, which are under common transcriptional control. In addition, statistical tools are offered to the user, which allow validation and comparison of those co-responses that were discovered by gene queries performed on the currently available set of pre-selectable datasets.

Availability: All co-response databases can be accessed through the CSB.DB Web server (http://csbdb.mpimp-golm.mpg.de/).

Contact: Steinhauser@mpimp-golm.mpg.de

INTRODUCTION

The availability of full genomic information (Goffeau *et al.*, 1996; Blattner *et al.*, 1997; The Arabidopsis Genome Initiative, 2001; Lander *et al.*, 2001) facilitated the development and spurred application of multi-parallel techniques to monitor the cellular inventory. Functional assignment of novel and partially characterized genes will continue to be the most important goal in biological science (Wu *et al.*, 2002; Shen-Orr *et al.*, 2002). Modern functional genomics encompasses technologies designed for the systematic investigation of gene function at all levels of a living cell, namely the genome, the transcriptome, the proteome and the metabolome (Fiehn *et al.*, 2000; Lockhart and Winzeler, 2000; Corbin *et al.*, 2003). The combined and multi-parallel analyses allow the investigation of complex biological processes at full systems level (Kitano, 2002) and may become the empirical basis of understanding the paradigm of life's complexity pyramid (Oltvai and Barabási, 2002). A future task will be the discovery of functional interaction within and among the levels of the cellular inventory, e.g. among metabolome and transcriptome (Urbanczyk-Wochniak *et al.*, 2003), and to extend knowledge from an organism-specific level towards general, organism-independent principles (Oltvai and Barabási, 2002). Hypotheses on units of genes with common function need to be associated with the currently available public knowledge of the complete cellular inventory. This information is made available in highly frequented but separate biological databases, which harbour genomic data (Mewes *et al.*, 2004), gene expression data (Sherlock *et al.*, 2001), information on protein properties (Schomburg *et al.*, 2004), metabolites and metabolic pathways (Kanehisa *et al.*, 2004). To gain insight into the functional organization of biological networks, specialized databases are required that are designed to store, handle, analyse and display the data derived from multi-parallel measurements. The comprehensive systems-biology database (CSB.DB) was developed to integrate biostatistical analyses on multi-parallel measurements of different organisms, such as *Escherichia coli*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. The present goal of CSB.DB is to present a publicly accessible resource for large-scale computational analyses on transcript co-response data, which mirror the large functional network of the cellular inventory and may serve as the basis for more sophisticated means of elucidating gene function.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

PROJECT OVERVIEW

The main focus of the CSB project is the generation of easily accessible knowledge about apparent gene-to-gene interactions in sets and subsets of publicly available transcript profiling data. We implicitly make the assumption that common transcriptional control of genes is reflected in co-responding, synchronous changes in transcript levels (Steinhauser *et al.*, 2004). For future implementations, we will extend this concept to the interaction of genes with other elements of the cellular inventory, such as metabolites (Urbanczyk-Wochniak *et al.*, 2003). Currently, no public convention exists as to which numerical approach is best applied to detect and validate the co-response of changing transcript levels. For this reason, we integrated a range of different statistical and computational algorithms, which are routinely applied in various research areas, such as Pearson's correlation, Kendall's correlation, Spearman's correlation, Euclidian distance and mutual information. Furthermore, we selected a range of different datasets, which comprises three organisms and were generated by different microarray technology platforms. Thus, the user of our co-response calculations is free to test the results on different datasets and species.

The basic aim of the CSB project is to supply researchers in the field of systems biology, molecular and applied biology with statistical tools to access transcriptional co-response. We concentrate on the validation of gene co-response without requirement for the user to have a priori knowledge about statistical methods and computational algorithms. We decided to preferentially facilitate access for those biologists who are interested in a specific gene of interest or small sets of genes. In this sense, our approach is similar to simple BLAST searches (Altschul *et al.*, 1990) of single or small number of genes. However, our approach towards the generation of novel functional hypotheses is based exclusively on simultaneous changes in transcript levels and does not require structural or sequence information.

IMPLEMENTATION AND STRUCTURE

CSB.DB is accessible via the Internet without the need to download special software to the client computer. The only system requirements are a JavaScript enabled recent web-browser and the ability to display PDF files. Only some advanced features require the JAVA extension. CSB.DB operates on a multiprocessor SuSE Linux (http://www.suse.de/de/index.html) system under an Apache web server (http://www.apache.org), and uses SAPDB (http://www.sapdb.org) as the database management system that stores the results of co-response analyses. CGI scripts, which connect the user queries with the database, are implemented in the PERL language (http://www.perl.com). The dynamic validation of discovered co-responses, graphical visualization as well as statistic algorithms, such as bootstrap and jack-knife analyses, are implemented as R (http://www.r-project.org) scripts, and can be invoked upon user selection to generate a PDF output. These files can be optionally downloaded by the user for further reference and documentation.

CSB.DB currently contains only co-response analyses, which are derived from publicly available expression profiling data. The calculated co-response data comprise pairwise gene correlations of three model organisms, namely *E.coli* (Steinhauser *et al.*, 2004), *S.cerevisiae* and *A.thaliana* (Fig. 1A).

DATABASES AND QUERIES

Co-response calculations based on changes in mRNA levels are the basis of functional annotation in CSB.DB and extend conventional predictions of gene function by analysis of gene homology (Wu *et al.*, 2002). Publicly available expression profiles of various organisms represent a rich resource for cross-experiment co-response analysis of genes, but need to be critically appraised. We used transcript profiles that were quality checked according to the recommendations of the respective technology platform. Furthermore, we included only accurately measured gene spots for the assembly into multi-conditional expression data matrices. For example, our data matrices comprise approximately 20–50 independent transcript profiling experiments and contain only 5% missing values per gene. Besides quality checking and reduction in missing data, we chose two general strategies for combining transcript datasets prior to correlation analysis. (1) We selected representative transcript profiles of as many different experimental conditions as possible. This approach allowed the search for general, constitutive gene-to-gene correlations in each organism. (2) If available we selected subsets of only those profiles, which were generated in a single set of biological experiments or under common biological conditions. These datasets allowed investigations of conditional changes in gene-to-gene co-response as compared to constitutive co-responses. Correlations were computed with the cCoRv1.0 software (Steinhauser *et al.*, unpublished data) and stored in organism specific co-response databases (Fig. 1A).

Rank-ordered tables of pairwise gene correlations according to the selected correlation measure can be obtained using the single gene query option and using a selection of predefined ranking strategies (sGQ; Fig. 1B). Similar to typical BLAST queries, sGQ allows to define a gene of interest and to retrieve all genes associated with co-response, if the gene of interest is represented among the set of quality checked genes. Moreover, the variant of sGQ made available for the *Arabidopsis* co-response databases allows to select filtering according to the functional categories, which were reported previously together with the visualization tool MapMan (Thimm *et al.*, 2004). The sGQ output (Fig. 1C) is presented as a HTML table, which contains the rank, the gene identifier of the co-responding gene, the correlation measure, the gene

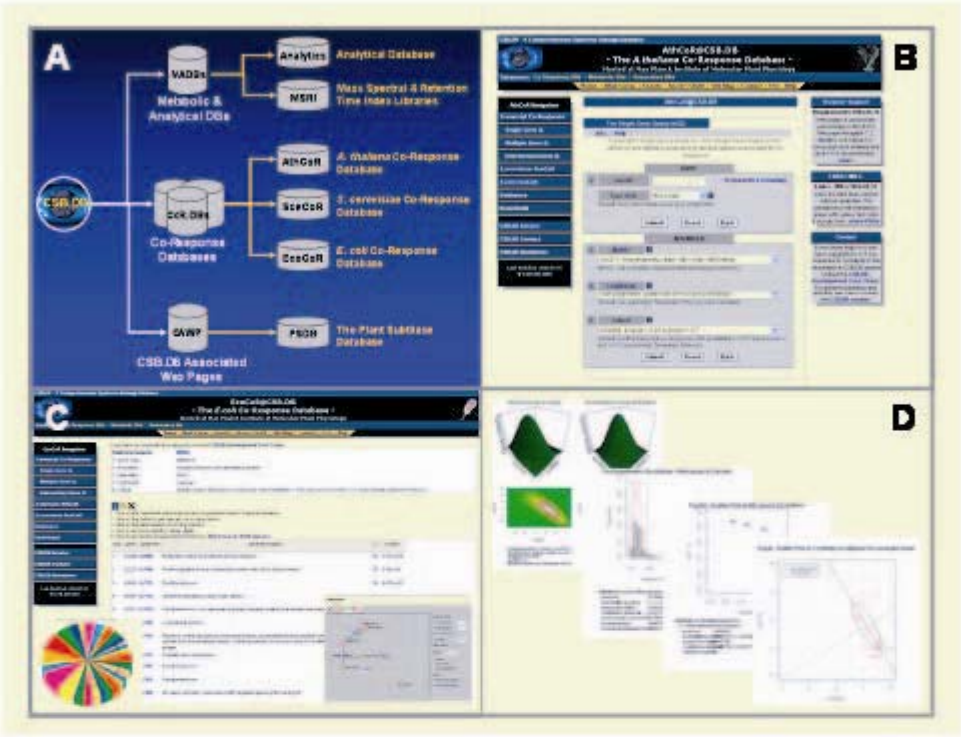


Fig. 1. Summarized overview of the current structure of CSB.DB (A) and selected examples of the available functionalities of the organism-specific co-response databases (B–D). (B) Represents one of the three possible query types, e.g. the single gene query sGQ in its HTML layout. The output of the queries is a HTML table, which contains in the case of *A.thaliana* a pie-chart summary on functional categories of the retrieved best ranking genes (C). (D) Shows examples of available gene-to-gene (bi-) plots, which can be invoked upon user request.

description, the number of pairs (n), the covariance (cov), the probability (*P*-value), the confidence interval (CI), the power, the mutual information [d(M), converted into to distance range] and the normalized Euclidean distance [d(E)]. These statistical parameters are dynamically calculated based on the underlying test distribution of the respective pre-selected correlation coefficient (Sokal and Rohlf, 1995; Bonett and Wright, 2000). Graphical summaries of the set of co-responding genes are based on various external functional classification efforts (Thimm *et al.*, 2004; Peterson *et al.*, 2001; Christie *et al.*, 2004) and/or the text search of the returned gene annotations (Fig. 1C). This survey of gene categories present in the hitlist is presented below the sGQ table.

Upon user request, a detailed statistical analysis may be obtained for a selected gene pair of interest. This additional validation on demand supports the detection of experimental outliers, which may be associated with technical errors or with the specific nature of a biological experiment. For this purpose, a variety of graphical plots are offered (Fig. 1D).

The multiple gene query option (mGQ) allows predefinition of up to 15 genes of interest and returns the complete set of available correlations among these genes. This option may be used to discover interdependences of genes, which are known to contribute to a common function or pathway. To visualize data, the interrelationship is also displayed as a co-response network with extensive filtering and layout options in JAVA enabled browsers (Fig. 1C).

Finally, an intersection gene query tool (isGQ) extracts those genes, which exhibit common correlations to at least two pre-defined genes of interest. The threshold settings, which are available for sGQ, may also be used for isGQ. The isGQ query may be used, if a few genes with a common

function are already known. Using the intersection mode that allows to find novel genes, which may be involved in this function, but cannot be discovered based on sequence homology.

OUTLOOK

We named CSB.DB 'A Comprehensive Systems-Biology Database', because we are convinced that the interpretation of gene co-response, which we currently make available to potential users, will in future require the integration of additional public resources on the present knowledge of the cellular inventory. Upon starting to use external functional classifications of genes, which among others include pathway and enzyme information, we implemented first access in our database to functional gene annotations. Thus, we laid the ground to retrieve biochemical reactions from publicly accessible metabolite databases starting from the result lists of highly correlated genes.

In addition, we previously described that the combined correlation analysis of changes in metabolite and mRNA levels may be highly informative and provide novel information (Urbanczyk-Wochniak *et al.*, 2003). Therefore, we will proceed to integrate profiling experiments and datasets into our database, which comprise measurements of changes in metabolite and transcript levels. Starting to use the same principles, which we apply to discover co-response in transcript datasets, we hope to unravel novel interactions between transcripts and metabolites. Thus, we are convinced that CSB.DB will develop into a highly useful and informative public resource.

ACKNOWLEDGEMENTS

We thank the staff of the SMD database (Sherlock *et al.*, 2001), the ASAP database (Glasner *et al.*, 2003) and the NASC Affymetrix Facility (Craigon *et al.*, 2004) for the establishment of public accessible resources of microarray data. We appreciate the work of all scientists, who submitted transcript profile data to these databases and thereby made comparative investigations possible. Furthermore, we thank the staff of the Free Software Foundation (FSF) for access to software under the terms of the GNU general public license. We are grateful to Prof. Lothar Willmitzer, Prof. Mark Stitt and the Max-Planck-Institute of Molecular Plant Physiology for support of the CSB project. Furthermore, the comments from Dr Leonard Krall, Dr Dirk Buessis and Stefan Kempa are gratefully acknowledged. The work of B.U. is partially financed by the GABI project 0312277D.

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignments search tool. *J. Mol. Biol.*, **215**, 403–410.

Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1462.

Christie,K.R., Weng,S., Balakrishnan,R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Feierbach,B., Fisk,D.G., Hirschman,J.E. *et al.* (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.*, **32**, D311–D314.

Corbin,R.W., Paliy,O., Yang,F., Shabanowitz,J., Platt,M., Lyons,C.E.,Jr, Root,K., McAuliffe,J., Jordan,M.I., Kustu,S., Soupene,E. and Hunt,D.F. (2003) Toward a protein profile of *Escherichia coli*: comparison to its transcription profile. *Proc. Natl Acad. Sci. USA*, **100**, 9232–9237.

Craigon,D.J., James,N., Okyere,J., Higgins,J., Jotham,J. and May,S. (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res.*, **32**, D575–D577.

Fiehn,O., Kopka,J., Dormann,P., Altmann,T., Trethewey,R.N. and Willmitzer,L. (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.*, **18**, 1157–1161.

Glasner,J.D., Liss,P., Plunkett,G.,III, Darling,A., Prasad,T., Rusch,M., Byrnes,A., Gilson,M., Biehl,B., Blattner,F.R. and Perna,J.T. (2003) ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res.*, **31**, 147–151.

Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 546–567.

Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.

Kitano,H. (2002) Systems biology: a brief overview. *Science*, **295**, 1662–1664.

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., Fitzthugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lockhart,D.J. and Winzeler,E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–836.

Mewes,H.W., Amid,C., Arnold,R., Frishman,D., Guldener,U., Mannhaupt,G., Munsterkotter,M., Pagel,P., Strack,N., Stumpflen,V., Warfsmann,J. and Ruepp,A. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, **32**, D41–D44.

Oltvai,Z.N. and Barabási,A.-L. (2002) Life's complexity pyramid. *Science*, **298**, 763–764.

Peterson,J.D., Umayam,L.A., Dickinson,T., Hickey,E.K. and White,O. (2001) The Comprehensive microbial resource. *Nucleic Acids Res.*, **29**, 123–125.

Schomburg,I., Chang,A., Ebeling,C., Gremse,M., Heldt,C., Huhn,G. and Schomburg,D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.*, **32**, D431–D433.

Shen-Orr,S.S., Milo,R., Mangan,S. and Alon,U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.*, **31**, 64–68.

Sherlock,G., Hernandez-Boussard,T., Kasarskis,A., Binkley,G., Matese,J.C., Dwight,S.S., Kaloper,M., Weng,S., Jin,H., Ball,C.A. *et al.* (2001) The Stanford Microarray Database. *Nucleic Acids Res.*, **29**, 152–155.

Sokal,R.R. and Rohlf,F.J. (1995) *Biometry: The Principles and Practice of Statistics in Biological Research*, 3rd edn. W.H. Freeman and Company, New York.

Steinhauser,D., Junker,B.H., Luedemann,A., Selbig,J. and Kopka,J. (2004) Hypothesis-driven approach to predict transcriptional units from gene expression data. *Bioinformatics*, **20**, 1928–1939.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Thimm,O., Blasing,O., Gibon,Y., Nagel,A., Meyer,S., Kruger,P., Selbig,J., Muller,L.A., Rhee,S.V. and Stitt,M. (2004) MAPMAN: a user-driven tool to display genomics datasets onto diagrams of metabolic pathways and other biological processes. *Plant J.*, **37**, 914–939.

Urbanczyk-Wochniak,E., Luedemann,A., Kopka,J., Selbig,J., Roessner-Tunali,U., Willmitzer,L. and Fernie,A.R. (2003) Parallel analysis of transcript and metabolic profiles: a new approach in systems biology. *EMBO Rep.*, **4**, 989–993.

Wu,L.F., Hughes,T.R., Davierwala,A.P., Robinson,M.D., Stoughton,R. and Altschuler,S.J. (2002) Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters. *Nat. Genet.*, **31**, 255–265.

Appendix D: Applications to Plant Environmental Stress Physiology

[17] Kaplan F, **Kopka J**, Haskell DW, Zhao W, Schiller KC, Gatzke N, Sung DY, Guy CL (2004) Exploring the temperature-stress metabolome of Arabidopsis. *Plant Physiology* 136 (4): 4159-4168
(http://dx.doi.org/10.1104/pp.104.052142)
(http://www.plantphysiol.org/cgi/content/abstract/136/4/4159)
Free article (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC535846/?tool=pubmed)

[18] Kaplan F, **Kopka J**, Sung DY, Zhao W, Popp M, Porat R, Guy CL (2007) Transcript and metabolite profiling during cold acclimation of Arabidopsis reveals an intricate relationship of cold-regulated gene expression with modifications in metabolite content. *Plant Journal* 50 (6): 967-981
(http://dx.doi.org/10.1111/j.1365-313X.2007.03100.x)
(http://www3.interscience.wiley.com/journal/118488580/abstract?CRETRY=1&SRETRY=0)

Co-supervision of Dr. Fatma Kaplan in a long-term cooperation with Prof. Dr. Charles L. Guy (University of Florida, Gainsville, USA) on metabolic aspects of the temperature stress acclimation response of Arabidopsis thaliana at transcriptome and metabolome levels. The GC-MS based metabolite profiling experiments were co-designed, the profiling performed and the data mining executed. In addition, the concept of comparing transcriptional and metabolomic effects for the discovery of systems level interactions which may control metabolic processes was contributed. The impact of this work has recently been reviewed by Guy CL et al. (2008b).

[19] Colebatch G, Desbrosses GG, Ott T, Krusell L, Montanari O, Kloska S, **Kopka J**, Udvardi MK (2004) Global changes in transcription orchestrate metabolic differentiation during symbiotic nitrogen fixation in *Lotus japonicus*. *Plant Journal* 39 (4): 487-512
(http://dx.doi.org/10.1111/j.1365-313X.2004.02150.x)
(http://www3.interscience.wiley.com/journal/118793989/abstract)

[20] Desbrosses GG, **Kopka J**, Udvardi MK (2005) *Lotus japonicus* metabolic profiling. Development of gas chromatography-mass spectrometry resources for the study of plant-microbe interactions. *Plant Physiology* 137 (4): 1302-1318
(http://dx.doi.org/10.1104/pp.104.054957)
(http://www.plantphysiol.org/cgi/content/abstract/137/4/1302)
Free article (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1088322/?tool=pubmed)

[21] Sanchez DH, Siahpoosh MR, Roessner U, Udvardi MK, **Kopka J** (2008) Plant metabolomics reveals conserved and divergent metabolic responses to salinity. *Physiologica Plantarum* 132 (2): 209-219
(http://dx.doi.org/10.1111/j.1399-3054.2007.00993.x)
(http://www3.interscience.wiley.com/journal/119395394/abstract)

[22] Sanchez DH, Lippold F, Redestig H, Hannah M, Erban A, Kraemer U, **Kopka J**, Udvardi MK (2008) Integrative functional genomics of salt acclimation in the model legume *Lotus japonicus*. *Plant Journal* 53 (6): 973-987
(http://dx.doi.org/10.1111/j.1365-313X.2007.03381.x)
(http://www3.interscience.wiley.com/journal/119410763/abstract)

[23] Sanchez DH, Redestig H, Kraemer U, Udvardi MK, **Kopka J** (2008) Metabolome-ionome-biomass interactions: What can we learn about salt stress by multiparallel phenotyping? Plant Signaling and Behavior 3 (8): 598-600
(PMID: 19704810, http://dx.doi.org/not available)
Free article (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2634509/)

These studies are result of a long-term equal cooperation with Dr. Michael K. Udvardi (Max Planck-Institute of Molecular Plant Physiology; since 2007, Nobel Foundation, Oklahoma, USA) on metabolic aspects of Lotus japonicus plant-microbe interactions and salt stress acclimation. My contributions started with co-supervision of Dr. Guilhem Desbrosses and proceeded to equal contribution and the main supervision Dr. Diego H. Sanchez. In the initial studies my laboratory enabled the metabolite profiling analyses, performed data mining of the resulting metabolite profiles for relevant results and contributed to the interpretation of metabolic versus transcriptional events. With the change to project leadership I now direct our studies towards enhanced and predictive experimental designs towards elucidation of the role of metabolism in *Lotus japonicus* salt stress acclimation.

Exploring the Temperature-Stress Metabolome of Arabidopsis[1][w]

Fatma Kaplan, Joachim Kopka, Dale W. Haskell, Wei Zhao, K. Cameron Schiller, Nicole Gatzke, Dong Yul Sung[2], and Charles L. Guy*

Plant Molecular and Cellular Biology Program, Environmental Horticulture (F.K., D.W.H., D.Y.S., C.L.G.), Department of Statistics (W.Z.), and Pharmacy Health Care Administration (K.C.S.), University of Florida, Gainesville, Florida 32611; and Max Planck Institute of Molecular Plant Physiology, 14476 Golm, Germany (J.K., N.G.)

Metabolic profiling analyses were performed to determine metabolite temporal dynamics associated with the induction of acquired thermotolerance in response to heat shock and acquired freezing tolerance in response to cold shock. Low-M_r polar metabolite analyses were performed using gas chromatography-mass spectrometry. Eighty-one identified metabolites and 416 unidentified mass spectral tags, characterized by retention time indices and specific mass fragments, were monitored. Cold shock influenced metabolism far more profoundly than heat shock. The steady-state pool sizes of 143 and 311 metabolites or mass spectral tags were altered in response to heat and cold shock, respectively. Comparison of heat- and cold-shock response patterns revealed that the majority of heat-shock responses were shared with cold-shock responses, a previously unknown relationship. Coordinate increases in the pool sizes of amino acids derived from pyruvate and oxaloacetate, polyamine precursors, and compatible solutes were observed during both heat and cold shock. In addition, many of the metabolites that showed increases in response to both heat and cold shock in this study were previously unlinked with temperature stress. This investigation provides new insight into the mechanisms of plant adaptation to thermal stress at the metabolite level, reveals relationships between heat- and cold-shock responses, and highlights the roles of known signaling molecules and protectants.

Environmental stresses arise from conditions that are unfavorable for the optimal growth and development of organisms (Levitt, 1972; Guy, 1999). Environmental stresses can be classified either as abiotic or biotic. Abiotic stresses are produced by inappropriate levels of physical components of the environment, including temperature extremes. Biotic stresses are caused by pathogens, parasites, predators, and other competing organisms. Even though biotic and abiotic stresses cause injury through unique mechanisms that result in specific responses, all forms of stress seem to elicit a common set of responses (Levitt, 1972). For instance, both biotic and abiotic stresses can result in oxidative stress through the formation of free radicals, which are highly destructive to lipids, nucleic acids, and proteins (Mittler, 2002). Another example is water stress, which is produced as a secondary stress by chilling, freezing, heat, and salt, as a tertiary stress by radiation, and, of course, as a primary stress during drought (Levitt, 1972).

The ability of most organisms to survive and recover from unfavorable conditions is a function of basal and acquired tolerance mechanisms. Acquired tolerance involves a set of mechanisms that can transiently extend or improve overall stress tolerance (Levitt, 1972; Hallberg et al., 1985; Guy, 1999; Thomashow, 1999) following exposure to moderate stress conditions. For example, if plants are preexposed to a nonlethal high temperature, they can acquire enhanced tolerance to otherwise lethal high temperatures. Similarly, many plants can tolerate a greater level of freezing stress when they are preexposed to nonlethal low temperatures. The ability to acquire enhanced tolerance to heat stress is known as acquired thermotolerance, while enhanced tolerance to freezing could be termed acquired freezing tolerance (Guy et al., 1985; Hallberg et al., 1985; Guy, 1999; Thomashow, 1999).

It has long been suspected and is now well accepted that temperature acclimation results from a complex process involving a number of physiological and biochemical changes, including changes in membrane structure and function, tissue water content, global gene expression, protein, lipid, and primary and secondary metabolite composition (Levitt, 1972; Gilmour et al., 2000; Shinozaki and Dennis, 2003). Recent advances in genome sequencing and global gene expression analysis techniques have further established the

[1] This work was supported by National Aeronautics and Space Administration (grant no. NAG10–316), by the U.S. Department of Agriculture (National Research Initiative grant nos. 2000–35100–9532 and 2002–35100–12110), and by the Institute of Food and Agricultural Sciences at the University of Florida. This article is Journal Series Number R–10483.

[2] Present address: Division of Biological Sciences, University of California, San Diego, CA 92093.

* Corresponding author; e-mail clguy@ufl.edu; fax 352–392–1413.

[w] The online version of this article contains Web-only data.

Article, publication date, and citation information can be found at www.plantphysiol.org/cgi/doi/10.1104/pp.104.052142.

multigenic quality of environmental stress responses and the complex nature of temperature acclimation (Seki et al., 2001; Fowler and Thomashow, 2002; Kreps et al., 2002). Literally hundreds of genes have been linked with environmental stress responses. By contrast, less is known about stress responses of plants at the metabolite and metabolome level (Cook et al., 2004; Rizhsky et al., 2004).

Global metabolite profiling analysis holds the promise to permit simultaneous monitoring of precursors, intermediates, and products of metabolic pathways. It is a discovery tool that can detect and monitor unidentified mass spectral tags (MSTs) as well as identified metabolites that play important roles in metabolism and physiology and, in the context of this work, stress tolerance. We have performed metabolite profiling analysis using gas chromatography-mass spectrometry (GC-MS) to determine similarities and differences in temporal metabolite responses and to identify novel compounds that exhibit temperature-specific responses during the induction of acquired thermotolerance in response to heat shock (HS), and during induction of acquired freezing tolerance in response to cold shock (CS). Metabolite profiling has revealed that CS influenced metabolism more profoundly than HS. However, the majority of HS responses were shared with CS, uncovering a novel relationship between HS and CS responses not previously known. This investigation provides a new viewpoint regarding metabolomic mechanisms of plant adaptation to thermal stress.

RESULTS

Temperature-Stress Acclimation Trends

Basal heat-stress tolerance for Arabidopsis (*Arabidopsis thaliana*) aerial tissues was between 43°C and 44°C, using an immersion assay that was chosen to minimize experimental variation due to the influence of transpirational leaf cooling (Gates, 1968). Upon transfer to an environment with an ambient air temperature of 40°C, Arabidopsis shoots began to undergo induction of acquired thermotolerance. Within 15 min of exposure to 40°C, shoot thermotolerance had increased by 1°C (Fig. 1A), and over a 4-h period tissue thermotolerance increased by as much as 5°C.

Basal freezing tolerance for Arabidopsis was −4°C when grown in a controlled environment at 20°C. Upon exposure to 4°C, freezing tolerance increased from −4° to −11°C over the course of 96 h (Fig. 1B). Enhanced freezing tolerance was observed as early as 6 h and continued to increase until 96 h of exposure. Freezing tolerance gradually diminished after 96 h. By contrast, plants returned to 20°C after 96 h of exposure to 4°C (Fig. 1B) underwent a process known as deacclimation (DA), leading to a significant decline in freezing tolerance. Approximately one-half of the induced freezing tolerance was lost within 24 h of return to 20°C.

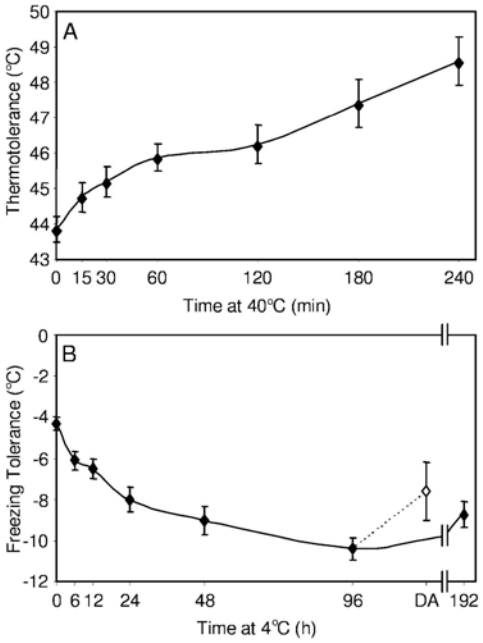


Figure 1. Kinetics of acquired thermotolerance and freezing tolerance induction in Arabidopsis. A, HS; B, CS. White symbol in B represents 24-h DA at 20°C after 96 h of CS. Error bars represent the 95% confidence interval of the mean. Electrolyte leakage assays were used to determine temperature causing lethal injury for acquired thermotolerance and acquired freezing tolerance.

Principal Component Analysis

Principal component analysis (PCA) was performed to test for the presence of differences between HS and CS, assess overall experimental variation, and determine individual time-point variation. PCA revealed that the four highest ranking components accounted for 61% of the total variance within the dataset (Fig. 2; Supplemental Table I, available at www.plantphysiol.org). Inspection of three of these components allowed consistent classification of the different treatment/time-point samples:

(1) Differential response to HS and CS. The first principal component (Fig. 2), accounting for 42.1% of the variance, indicated a strong differential response of HS and CS at the metabolic level.
(2) Differential response with respect to the time series. The second component (Fig. 2), accounting for 8.5% of total variance, resolved the time series of both HS and CS responses. The heat response followed a continuous linear trend, whereas samples of the cold-response time series were arranged in a bipartite but continuous sequence, indicative of continuous transient changes during cold

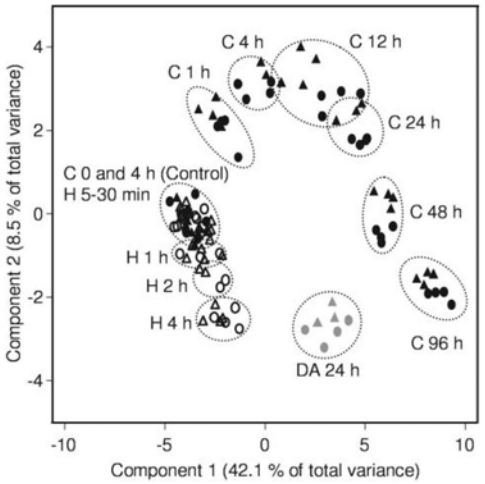


Figure 2. Principal component analysis. Component 1 (differential response to HS and CS) and Component 2 (differential response with respect to the time series) are plotted on the axes. All samples of this investigation are represented. CS samples (black), HS samples (white), untreated control samples (black), DA samples (gray), first experiment (triangles), and second experiment (circles) are shown. Time points are indicated within the graph.

acclimation. Early cold responses were opposite to the heat response, whereas late cold and heat responses were colinear.

(3) Differential response of DA. The third component, accounting for 5.9% of total variance (Supplemental Table I), separated deacclimated samples from all other samples and clearly established their distinct metabolic phenotype compared to controls.

(4) Component 4 and all subsequent components did not provide any further differentiation between sample types.

PCA analysis showed that the temperature treatment and time series effects clearly contributed most to the total variance within the data set. By contrast, the within-time point variance was low, as only slight shifts within the time sequence were observed (Fig. 2). The negligible interexperimental variation demonstrates the robustness of the experimental design.

Temporal Alterations of Metabolite Content in Response to Temperature Shock

We investigated sustained and transient changes with respect to three major categories of temporal response: early, intermediate, and late. Statistical analysis was performed on known metabolites and MSTs. Metabolites and MSTs were screened for significant changes ($P < 0.05$) in at least one time point after either heat or cold treatment. Of the 497 low-M_r polar compounds detected, the levels of 143 were altered in response to HS (Supplemental Table II), and the levels of 311 were changed in response to CS (Supplemental Table III).

Out of the 143 HS-responsive metabolites and MSTs, 85 showed a sustained (Fig. 3, A, C, E, and F) or transient (Fig. 3, B and D) increase or decrease (Supplemental Table II). The majority of the metabolite responses to high temperature occurred within the first 30 min, when thermotolerance was increasing (Fig. 1A). A total of 58 metabolites and MSTs showed an early, 9 showed an intermediate, and 18 showed a late increase or decrease in response to HS. Components of amino acid and carbohydrate metabolism were affected by HS. Coordinate increases in the pool sizes of a number of amino acids (Asn, Leu, Ile, Thr, Ala, Leu, and Val) derived from oxaloacetate and pyruvate were observed (Fig. 3A). Not surprisingly, fumarate and malate (oxaloacetate precursors) contents were similarly increased. Also, a small group of

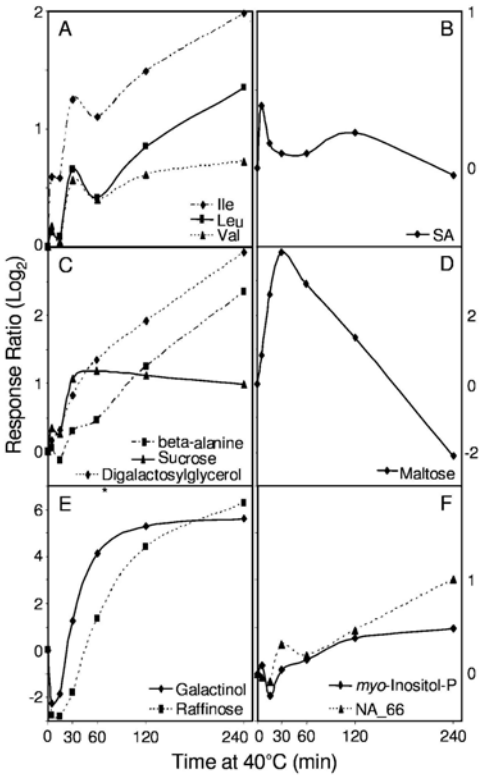


Figure 3. Representative HS metabolite responses. A to D demonstrate early; E, intermediate; F, late sustained and transient increase patterns. NA followed by number is an unidentified MST. If a tentative identification is available, MST is characterized with an asterisk.

amine-containing metabolites (β-Ala, 4-aminobutyric acid [GABA], and putrescine) with protective properties appeared to be coordinately increased. Further, a select group of well-known carbohydrates were affected (Fig. 3, C, D, E, and F), such as maltose, Suc, raffinose, its precursors galactinol, myoinositol, and cell-wall monosaccharides.

In contrast with HS, alterations in metabolite and MST contents were evenly distributed across all temporal stages of CS (Fig. 1B). Out of 311 CS-responsive compounds, 229 showed a clear sustained (Fig. 4, A, C, E, and G) or transient (Fig. 4, B, D, F, and H) increase or decrease (Supplemental Table III). The pool sizes of 92 metabolites and MSTs showed an early, 66 showed an intermediate, and 71 showed a late increase or decrease to CS. Overall, amino acids, TCA cycle intermediates, and many metabolites of carbohydrate metabolism were affected by CS. Parallel to HS, coordinate increases in the pool sizes of amino acids derived from oxaloacetate and pyruvate were observed during CS. Coordinate increases in the pool sizes of aromatic amino acids (Trp, Phe, and Tyr) were followed by increased pool sizes of phenylpropanoid pathway intermediates (cis-ferulic, cis-sinapic, and trans-sinapic acid). In addition, the pool sizes of amino acids (Pro, Arg, Cys, Gly, and Ser) derived from α-ketoglutarate and from 3-phosphoglycerate were also increased. Particularly during CS, the pool sizes of most TCA cycle intermediates, as were early glycolytic intermediates, were increased. Regarding the latter, there was a clear and profound shift in hexose metabolism that linked with di- and trisaccharide accumulations (Glc, Fru, Glc-6-P, Fru-6-P, myoinositol-P, Man-6-P, galactinol, Suc, and raffinose).

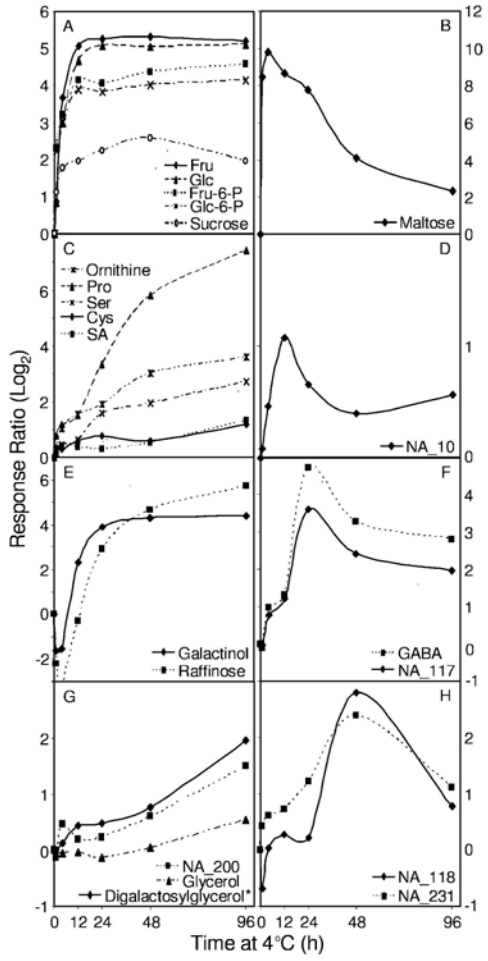


Figure 4. Representative CS metabolite responses. A to C demonstrate early; D to F, intermediate; G and H, late sustained and transient increase patterns. NA followed by number is an unidentified MST. If a tentative identification is available, MST is characterized with an asterisk.

Specific Temperature-Shock Responses

In order to determine similarities and differences between HS and CS responses, individual metabolites and MST profiles were compared and contrasted. Metabolites and MSTs exhibiting heat-specific (4%), cold-specific (38%), DA-specific (2%), and both HS and CS responses (25%) were identified (Fig. 5). About 31% of the metabolites and MSTs did not respond to either form of temperature shock.

Metabolites and MSTs that showed altered levels during HS, but not to CS, were considered HS specific. Eighteen compounds appeared to be heat specific (Table I), and three were identified (uracil, citramalate, and quinic acid).

Metabolites and MSTs that showed altered concentrations to CS but did not show significant changes during HS were considered CS specific. Of the 311 metabolites that responded to CS, the majority (186) was not responsive to HS (Table I). Of the 186 CS-specific metabolites, the levels of 140 increased, while 46 decreased. CS-specific metabolites included aromatic amino acids (Phe and Trp), intermediates in the phenylpropanoid pathway, α-ketoglutarate, 3-phosphpoglycerate derivative amino acids, and some of the early intermediates of the glycolytic pathway.

The levels of 12 MSTs increased in response to DA but were not altered in response to HS or CS, suggesting their direct involvement in the recovery process from long-term cold stress.

A total of 125 metabolite levels were altered in response to both HS and CS, 32 exhibited a differential and 93 exhibited a common response. Of the 32, the levels of 7 metabolites decreased during HS (Table I) but increased during CS. The levels of the remaining 25 metabolites (Table I) increased when exposed to HS but

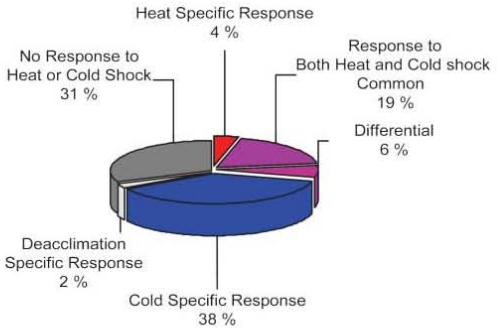


Figure 5. Proportionality of metabolite responses to temperature shock. A total of 497 metabolites and MSTs were detected by GC-MS.

decreased during CS. The 93 metabolites and MSTs that exhibited a common response to HS and CS represented two-thirds of the heat-responsive and one-third of the cold-responsive metabolites. The levels of 42 metabolites and 44 MSTs increased, while 7 decreased (Table I). Contents of oxaloacetate and pyruvate derivative amino acids, polyamines, and carbohydrates were increased under both HS and CS. A number of the metabolites in this group either have compatible solute properties or serve as precursors for secondary metabolites that protect plants against pathogens.

Common Stress Metabolites

In order to determine whether common temperature-response metabolites may play a role in overall environmental stress tolerance, the 42 metabolites with known identity were compared to those in the published literature for abiotic and biotic stress (Table I). As expected, many of them were reported to increase in response to other environmental stresses (shown with asterisk in Table I). Examples include salicylic acid (SA), GABA, Tyr, Leu, Val, Suc, and maltose (Srivastava et al., 1980; Handa et al., 1983; Mayer et al., 1990; Metraux et al., 1990; Malamy et al., 1990; Fan et al., 1993; Schmelz et al., 2003; Rizhsky et al., 2004).

The levels of the remaining 14 metabolites (Table I) were not previously reported to increase in response to temperature stress or any other environmental stresses. This grouping of 14 represents a metabolite discovery approach in the context of establishing a linkage in temperature-stress responses. The veracity of the linkage of these 14 metabolites with temperature stress was validated by a recent study (Cook et al., 2004) linking 9 of the metabolites with long-term cold acclimation of Arabidopsis, while this study was in review.

Statistical Linkage of Key Metabolites in the Metabolite Profile with Acquired Tolerances

In order to further investigate the contributors of the components of PCA, the metabolite loadings in components 1, 2, 3, and 4 were compared with the ANOVA results. Based on ANOVA, the 19 metabolites and MSTs that showed quantitative signal increases in component 1 mainly consisted of metabolites that increased in response to CS. The signature metabolites for component 1 were Glu, Pro, Arg, Fru-6-P, and MSTs ([NA_154], [949; glucopyranose], [861; glucopyranose], and [539; Phe]). Component 2 was mainly composed of metabolites that showed common or differential response to both HS and CS, or cold-specific metabolites. The signature metabolites for the common response were maltose, galactinol, and raffinose; for the differential response glc, glc-6-P, and MSTs ([NA_1] and [852; aminomalonic acid]); and for the cold-specific response Pro, Fru-6-P, and MSTs ([612; Pro], [NA_84], and [NA_154]). Component 3 was mainly composed of metabolites that quantitatively increased in response to CS and decreased to control levels when deacclimated for 24 h. The signature metabolites here were galactinol, raffinose, and MSTs ([NA_154] and [497; gluconic acid 1,4 lactone]). Component 4 largely extracted metabolites that did not change quantitatively in response to either HS or CS, and they were all MSTs ([NA_159], [NA_264], [NA_267], [NA_271], [NA_302], [NA_335], [NA_341], [NA_359], and [674; Gln]). Taken together, these metabolites and MSTs of components 1 to 3 are likely to play either a direct role in essential mechanisms of acquired tolerances or an indirect role as a consequence of occupying a central role in some aspects of cell metabolism.

DISCUSSION

Metabolites have a number of functions in addition to those of intermediary metabolism. They act as signaling/regulatory agents, compatible solutes, antioxidants, or in defense against pathogens. Our results provide new insight into mechanisms of plant adaptation to thermal stress at the metabolite level, highlight the roles of known signaling molecules and protectants, and reveal a previously unrecognized interrelationship of HS and CS responses.

Plants have several well-known regulatory metabolites that function in a number of plant growth and development processes. Some are also involved in plant environmental stress processes (Klee, 2003). In this study, SA levels showed a very rapid transient increase in response to HS at 5 min (Fig. 3B). Similarly, SA levels became elevated during CS, exhibiting a biphasic response starting at 1 h, peaking at 4 and 12 h, decreasing at 24 h, and then continuously increasing to 96 h (Fig. 4C). These findings firmly implicate SA as an early signaling molecule in temperature-stress responses. This is important for a number of reasons. SA has a key role in systemic acquired resistance to pathogens such as bacteria, fungi, and viruses (Metraux et al., 1990; Schmelz et al., 2003), and increases in SA levels are positively correlated with the

Table I. *Influence of temperature shock on metabolite levels*

I, Increase; D, decrease; N, no significant change in metabolite concentration. Names in brackets precede a match value for an unidentified compound. These names indicate best mass spectral similarity on a scale of 0 to 1,000 (1,000 is identical) to the indicated compound. *, Metabolite showed increase in other environmental stress conditions.

HS-Specific Response			Differential Response		
Metabolites	HS	CS	Metabolites	HS	CS
Uracil	I	N	Phosphoric acid	D	I
D-(−)-Quinic acid	I	N	Glc	D	I
13 MSTs	I	N	Glc-6-P	D	I
Citramalic acid	D	N	[798; Fru]	D	I
3 MSTs	D	N	3 MSTs	D	I
CS-Specific Response			**Common Response**		
Allantoin	N	I	2-Ketoglutaric acid	I	I
cis-Aconitic acid	N	I	β-Ala*	I	I
cis-Ferulic acid	N	I	Citric acid*	I	I
cis-Sinapic acid	N	I	Erythritol	I	I
Fru-6-P	N	I	Erythronic acid	I	I
GlcUA	N	I	Fru*	I	I
Glyceric acid-3-P	N	I	Fumaric acid	I	I
L-(+)-Ascorbic acid*	N	I	GABA*	I	I
L-Arg*	N	I	Galactinol*	I	I
L-Cys*	N	I	Galactonic acid	I	I
L-Glu*	N	I	Glycerol*	I	I
L-Gln*	N	I	Gly*	I	I
L-Phe*	N	I	L-Ala*	I	I
L-Pro*	N	I	L-Asn*	I	I
L-Ser*	N	I	L-Glycerol-3-P	I	I
L-Trp*	N	I	L-HomoSer*	I	I
Maleic acid	N	I	L-Ile*	I	I
Man-6-P	N	I	L-Leu*	I	I
Norvaline	N	I	L-Lys*	I	I
O-Acetyl-L-Ser	N	I	L-Met*	I	I
Octadecanoic acid	N	I	L-Thr*	I	I
PyroGlu	N	I	L-Tyr*	I	I
Sorbitol*	N	I	L-Valine*	I	I
trans-Sinapic acid	N	I	Malic acid*	I	I
[497; Gluconic acid-1,4-lactone]	N	I	Maltose*	I	I
[529; Indole-3-acetic acid]	N	I	Melibiose*	I	I
[539; Phe]	N	I	Myoinositol-P	I	I
[612; Pro]	N	I	Orn	I	I
[614; Gln	N	I	Putrescine*	I	I
[640; Putrescine]	N	I	Raffinose*	I	I
[734; L-Asp]	N	I	Ribose	I	I
[861; Glucopyranose]	N	I	SA*	I	I
[889; 1,6-AnhydroGlc]	N	I	Succinic acid	I	I
[949; Glucopyranose]	N	I	Suc*	I	I
106 MSTs	N	I	Threonic acid	I	I
Isocitric acid	N	D	Threonic acid-1,4-lactone	I	I
Lactic acid	N	D	Trehalose*	I	I
44 MSTs	N	D	Tyramine	I	I
Differential Response			Xyl	I	I
Ara	I	D	[721; Glucaric acid]	I	I
Man	I	D	[732; Pipecolic acid]	I	I
myo-inositol	I	D	[861; Digalactosylglycerol]	I	I
Shikimic acid	I	D	44 MSTs	I	I
[852; Aminomalonic acid]	I	D	Dehydroascorbic acid dimer	D	D
[708; Ribonic acid]	I	D	Glyceric acid	D	D
[846; Xylitol]	I	D	L-Asp	D	D
18 MSTs	I	D	4 MSTs	D	D

level of resistance to pathogens in plants (Heil and Bostock, 2002). Accordingly, a recent study has linked SA with basal thermotolerance (Clarke et al., 2004), and exogenous application of SA or acetyl salicylate has been shown to enhance thermotolerance (Dat et al., 1998; Lopez-Delgado et al., 1998; Senaratna et al., 2000; Clarke et al., 2004). One study has shown endogenous SA levels to be elevated at 30 min after the onset of HS (Dat et al., 1998). Our findings place the increase in SA levels to within 5 min of the onset of HS, which strongly implies a role for SA in early HS-signaling and acquired thermotolerance. Further, increases in SA levels during temperature shock in this study and in other abiotic (drought and salt stress) and biotic stresses (Garcia et al., 1997; Metraux et al., 1990; Munne-Bosch and Penuelas, 2003; Schmelz et al., 2003) suggest that SA could be a key signal molecule in the initiation of plant tolerance to a variety of environmental stresses. It is logical that integrating SA signaling in temperature-shock responses could help plants prepare to defend themselves against pathogens when plant host-pathogen defense systems are weakened by environmental stress.

Many metabolites can act in defense mechanisms against pests such as insects, pathogenic fungi, and bacteria. These metabolites are generally derived from secondary metabolism, such as the phenylpropanoid, isoprenoid, alkaloid, or fatty acid/polyketide pathways (Dixon, 2001). However, precursors of these defense compounds emanate from primary metabolism. For example, branched-chain amino acids (Ile, Leu, and Val) serve as precursors for cyanogenic glycosides (Vetter, 2000). Aromatic amino acids (Trp, Phe, and Tyr) serve as precursors for indole glucosinolates, phytoalexins, alkaloids, lignins, flavonoids, isoflavonoids, and hydroxycinnamic acids (Dixon, 2001). In this study, increased levels of Ile, Leu, Val, and Tyr in response to both HS and CS (Table I; Fig. 3A) and increased levels of Trp and Phe in response to CS were observed (Table I). The increase in Ile, Leu, Val, and Tyr content in response to other abiotic and biotic stresses and heat is well known (Srivastava et al., 1980; Mayer et al., 1990; Rizhsky et al., 2004). Therefore, it is reasonable that one purpose for branched-chain amino acid accumulation is to support increased production of secondary metabolites as part of a defense response against pathogens during temperature stress. Such a response could constitute a preemptive defense against opportunistic attack by a pathogen on a stress-weakened host.

Metabolites of primary metabolism can act as signal molecules. A well-known example is Suc (Koch, 1996; Chiou and Bush, 1998; Roitsch, 1999; Smeekens, 2000; Rolland et al., 2002; Moore et al., 2003), whose content in response to HS (within 5 min) and CS rose very rapidly (within 1 h) and was maintained throughout HS and CS exposures (Figs. 3B and 4A). Thus, Suc could be a candidate signaling molecule for both HS and CS, based on its early accumulation in response to temperature shock. Consistent with this notion, parallel microarray analysis has revealed that the promoters of a number of genes induced by both HS and CS contain sugar-responsive elements (F. Kaplan, D.Y. Sung, and C.L. Guy, unpublished data). Thus, a reasonable hypothesis that sugar signaling may be important in the establishment and maintenance of both acquired thermotolerance and freezing tolerance is worthy of further study. Arabidopsis knockouts defective in sugar signaling might prove valuable experimental tools in dissection of the signaling aspects of sugars during acquired thermotolerance and freezing tolerance. In addition to signaling role of Suc, the role of Suc and other soluble sugars (maltose, Glc, and Fru) as compatible solutes are well established during abiotic stresses, such as cold, drought, dessication, salt, and osmotic stress (Guy et al., 1992; Fan et al., 1993; Uemura et al., 2003; Rizhsky et al., 2004).

These findings support the notion that a multiplicity of primary metabolites could act collectively as compatible solutes. Compatible solutes (osmolytes, osmoprotectants) are low-M_r organic molecules that accumulate under stress conditions, and are considered to stabilize proteins and membranes and contribute to cell osmotic pressure. There are three general types of osmoprotectants: amino acids, quaternary ammonium compounds, and polyols (Bowlus and Somero, 1979; Yancey et al., 1982; Shahjee et al., 2002). During the onset of acquired thermotolerance, the content of Ala, Asn, β-Ala, Fru, GABA, glycerol, malate, maltose, Man, putrescine, raffinose, succinate, Suc, and trehalose increased in response to HS. A complementary GC-MS study consistent with our study also showed that 6 h of HS resulted in moderate increases in the content of β-Ala, glycerol, maltose, Suc, and trehalose (Rizhsky et al., 2004). The content of these metabolites increased with the persistence of the HS, and their proportions to each other changed as the duration of high-temperature exposure progressed (Fig. 3). Similar to the HS response, we observed increases in the content of many metabolites with known compatible solute properties during the development of acquired freezing tolerance at low temperature (Fig. 4). Examples include Ala, β-ala, Gly, Pro, Ser, Orn, putrescine, Fru, Glc, malate, maltose, and Suc. Persistence of low-temperature exposure led to increased quantities of these compatible solute metabolites and produced the accumulation of even more metabolites with compatible solute-like properties, notably Asn, GABA, glycerol, raffinose, sorbitol, succinate, and trehalose. This overall profile of metabolites with compatible solute-like properties suggests that it is a combination of compatible solutes that exerts additive or synergistic effects during the cold acclimation process and during the induction of thermotolerance. For instance, the Arabidopsis mutant (eskimo1), containing high levels of soluble sugars and Pro, possesses enhanced freezing tolerance (Xin and Browse, 1998). By contrast, failure to accumulate Suc and Glc results in reduced freezing tolerance in the Arabidopsis mutant (*sfr4*; McKown et al., 1996). The

enhanced freeze sensitivity of *sfr4* was shown to be due to the loss of osmotic responsiveness (LOR) of the protoplast. When exogenous Suc was supplied in vitro, LOR was reduced and freeze tolerance was improved in *sfr4* protoplasts (Uemura et al., 2003). The present metabolite profiling results implicate a more dynamic and larger compatible solute-like network then previously recognized. This may explain why attempts to engineer overproduction of a single compatible solute compound have not produced plants with high levels of stress tolerance (Chen and Murata, 2002).

The major advantage of metabolite profiling using a time-course design is that it permits simultaneous monitoring of entire metabolic pathways (precursors, intermediates, and products) and can reveal the subtle interplay of functionally related metabolites. In this study, a clear example of substrate and product relationship can be seen in Figures 3E and 4E. Galactinol, along with Suc, is an immediate precursor of raffinose, whose biosynthesis rate largely depends on the availability of Suc, galactinol, and the enzyme raffinose synthase (Taji et al., 2002). An increase in galactinol abundance clearly precedes the increase in raffinose content during both HS and CS, exactly as would be predicted by a classic substrate-product relationship. Such parallel relationships can be very powerful as a metabolite discovery tool that detects and monitors MSTs that may play important roles in stress adaptations.

In conclusion, the comparative metabolomic analysis of temperature-stress response has highlighted the roles of signaling molecules and implicated the action of a compatible solute network in temperature-stress tolerances. With respect to low-M_r polar compounds, CS, in a quantitative sense, influenced metabolism more profoundly than HS. The majority of metabolites responsive to CS were specific to CS. By contrast, a very large proportion of the HS metabolite response (about two-thirds) seemed to be shared with that of CS. Only a very small proportion of heat-responsive metabolites were heat specific. The present results support a number of paradoxical early observations that some cold-hardened plants were also more heat-stress tolerant (Alexandrov, 1964; Levitt, 1972). These results may also explain why HS seems to improve chilling tolerance (Lurie and Klein, 1991; McCollum et al., 1995; Saltveit, 2002; Saltveit and Hepler, 2004; Saltveit et al., 2004) in a number of cold-sensitive species. Therefore, treatment by HS for short periods of time might improve tolerance to acute HS and CS or induce an overall environmental stress tolerance. Additionally, heat- and cold-shocked plants increased concentrations of a number of metabolites that responded to a variety of environmental stimuli. The majority of the common temperature shock-responsive metabolites was accumulated during all stages of the development of acquired tolerances and did not seem to be specific to one particular phase in the development of acquired tolerance. Taken together, this work identifies a large number of potential metabolic targets for further in-depth investigations of acquired tolerances to temperature stress.

MATERIAL AND METHODS

Plant Growth

Arabidopsis (*Arabidopsis thaliana*; ecotype Columbia) plants were grown as described by Sung and Guy (2003) for 3 weeks. Plants were grown at 20°C with a photoperiod of 15/9-h light/dark cycle in growth cabinets for 3 weeks. Irradiance was provided by incandescent bulbs and cool-white fluorescent tubes and ranged between 100 and 150 μmol m^{-2} s^{-1} at canopy height. Plants were at an eight-leaf stage of development when the experiments were begun (Supplemental Fig. 1).

Electrolyte Leakage Assay for Thermotolerance

To measure acquired thermotolerance, plants were given a HS (40°C) for 0, 15, 30, 60, 120, 180, and 240 min beginning 2 h after the onset of the light period. At the indicated times, plants were immersed to the soil lines in 42°C, 44°C, 46°C, 48°C, and 50°C water for 10 min, and electrolyte leakage of the aerial portion of a plant was measured 3 d after heat treatment (Sung and Guy, 2003). For each time point and temperature, 14 independent experiments with 3 replications were done, except for 180 and 240 min. For time points 180 and 240 min, one experiment with 5 replications was done.

For measuring freezing stress, plants were given a CS (4°C) for 0, 6, 12, 24, 48, 96, and 192 h beginning 2 h after the onset of the light period. Also, plants were deacclimated for 24 h at 20°C after 96 h of CS. At the indicated times, plants were rapidly harvested, wrapped in water-saturated tissue paper, placed in a test tube, then placed in a controlled-temperature bath (Forma Scientific model 2425; Marietta, OH) and equilibrated for 30 min at 0°C. Then a chip of ice was placed in contact with the tissue paper and the temperature was lowered at a rate of 2°C h^{-1}. Tubes were removed at 1° intervals, placed on ice, and allowed to thaw overnight at 4°C. For each time point and temperature, three independent experiments with five replications were done. For DA, one experiment with five replications was done.

Electrolyte leakage of the aerial portions of the plants was measured according to Sung and Guy (2003). Aerial portions of plants were placed in scintillation vials containing 10 mL of distilled water and shaken for 1 h. After the first conductivity reading was made, the tissue was boiled for 2 min by microwave irradiation. After cooling to room temperature, a second conductivity reading was taken following shaking for a second hour. Relative electrolyte leakage was determined from the ratio of first and second conductivity measurements.

Metabolite Profiling

Time points for metabolite profiling during temperature-shock treatments were selected based on thermo- and freeze-tolerance time-course experiments. Temperature-shock treatments were initiated 2 h after the onset of the light period, which allowed for the harvest of all samples within the light period. Three-week-old 20°C-grown plants were placed at 40°C and sampled at 5, 15, 30, 60, 120, and 240 min of HS. At the same time, a second set of plants was placed at 4°C and sampled at 1, 4, 12, 24, 48, and 96 h of CS. After 96 h at 4°C, plants were returned to 20°C and sampled after 24 h. Additionally, untreated controls were taken at zero time of the experiment and 4 h after the experiment began. All samples were rapidly harvested, flash-frozen in liquid nitrogen (<30 s), and stored at −80°C until metabolite extraction. Two sets of temperature-stress experiments were performed, each comprising two to four replicate measurements per time point. Aerial tissues were ground in liquid nitrogen with pestle and mortar. Aliquots of 60 mg of frozen powder were extracted with hot MeOH/$CHCl_3$ and the fraction of polar metabolites processed as described (Wagner et al., 2003). Ribitol, isoascorbic acid, and deuterated Ala were added as internal standards. A C_{12}, C_{15}, C_{19}, C_{22}, C_{28}, C_{32}, and C_{36} *n*-alkane mixture was used for the determination of retention time indices (RI). Metabolite samples were derivatized as described by Fiehn et al. (2000) and Roessner et al. (2000), analyzed by an MD 800 GC-MS system

(ThermoQuest, Manchester, UK). Chromatograms were processed using the find algorithm of the MassLab version 1.4 software (ThermoQuest).

Metabolite Identification

GC-MS-based metabolite profiling detects and quantifies specific mass spectral fragments in defined retention time windows. Identification of these fragments was performed through standard addition experiments using pure authenticated compounds to confirm identity by retention time index and mass spectrum. Compounds were designated as metabolites if they were identified with a match >750 on a scale of 0 to 1,000 and RI deviation <3.0. Other unidentified compounds are designated as MSTs designated by the code NA and a unique number. In cases of high mass spectral similarity of MSTs to available commercial or custom mass spectral libraries, MSTs were named in square brackets by a preceding match value and a compound name taken from these libraries. Representative mass spectra and RI, which serve for metabolite identification in Arabidopsis, and novel identifications post-publication will be available through CSBDB (http://csbdb.mpimp-golm.mpg.de/csbdb/dbma/msri.html).

Statistical Analysis

PCA was performed with the S-Plus 2000 software package standard edition release 3 (Insightful, Berlin) on $\log_{10}$-transformed relative responses, $\log_{10}(R_i)$. Missing data were replaced with 0 for PCA. The denominator of the quotient, R_i, was the average response of nontreated control samples at zero time of the respective stress experiments ($R_i = N_i \times \text{avg}N_{t0}^{-1}$). Responses ($N_i$) were volume corrected for error during sample preparation or GC injection and normalized by the fresh weight of each sample.

One-way ANOVA was done using the Kruskal-Wallis test on metabolite response values (N_i). Nonparametric approach was chosen because it did not require normally distributed data, and it was also more resistant to the outliers in the data set that might lead to high fold changes. Changes in metabolite content with $P < 0.05$ were considered to be significant. Pair-wise comparisons between different treatments and time points were done using the Kruskal-Wallis test.

Classification Criteria of Metabolite Responses

Criteria for HS metabolic responses were as follows: in the early sustained response (0–4 h), a statistically significant change in metabolite levels as compared to zero-time control occurred at 5, 15, or 30 min and was maintained until 4 h. In the intermediate-sustained response (1–4 h), a statistically significant response occurred at either 1 or 2 h, was maintained until 4 h, but did not exhibit a significant response at 5, 15, and 30 min. In the late response (4 h), a statistically significant response occurred at 4 h, but no significant response was observed at 5, 15, 30, 60, and 120 min. In the transient response, compounds exhibited a statistically significant response when compared to zero-time control, 4-h diurnal control, and 4-h HS. Transient changes occurring at 5, 15, and 30 min were considered early transient, and those at 1 and 2 h were considered intermediate transient. Additionally, a 4-h untreated control was included in the analysis to filter diurnal responses from the HS data set.

The criteria for the CS metabolic responses were adjusted based on the Arrhenius equation relationship for respiratory processes (Yelenosky and Guy, 1977) to reflect the slowed metabolism during CS as follows: in the early sustained response (0–96 h), a statistically significant change in metabolite levels as compared to zero time occurred at 1 h or 4 h and was maintained until 96 h. In the intermediate-sustained response (12–96 h), a significant response occurred at 12 or 24 h, was maintained until 96 h, but no significant response occurred at 1 and 4 h. In the late response (48–96 h), a significant response occurred at either 48 or 96 h, but no significant response was observed at 1, 4, 12, and 24 h. In the transient response, metabolites showed a statistically significant response when compared to zero-time control and 96-h CS. Transient changes occurring at 1 and 4 h were considered early transient, those at 12 and 24 h were considered intermediate transient, and those at 48 h were considered late transient. Diurnally regulated metabolites were filtered from the 4-h early transient response group using the 4-h diurnal (untreated) control. Metabolites that did not fit the above criteria were classified as increase or decrease if they showed a significant increase or decrease at any time point during temperature shock when compared to zero-time control.

ACKNOWLEDGMENTS

We thank Michael Popp and Kil-Jae Lee for their help with this work, and B. Rathinasabathi, K.C. Cline, M.F. Thomashow, and Lonnie Ingram for critical reading of the manuscript. We also thank Lothar Willmitzer and Max Planck Society for continuing support.

Received August 23, 2004; returned for revision October 5, 2004; accepted October 5, 2004.

LITERATURE CITED

Alexandrov VY (1964) Cytophysiological and cytoecological investigations of heat resistance of plant cells toward the action of high and low temperature. Q Rev Biol **39:** 35–77

Bowlus RD, Somero GN (1979) Solute compatibility with enzyme function and structure: rationales for the selection of osmotic agents and end-products of anaerobic metabolism in marine invertebrates. J Exp Zool **208:** 137–151

Chen THH, Murata N (2002) Enhancement of tolerance of abiotic stress by metabolic engineering of betaines and other compatible solutes. Curr Opin Plant Biol **5:** 250–257

Chiou T-Z, Bush DR (1998) Sucrose is a signal molecule in assimilate partitioning. Proc Natl Acad Sci USA **95:** 4784–4788

Clarke SM, Mur LA, Wood JE, Scott IM (2004) Salicylic acid dependent signaling promotes basal thermotolerance but is not essential for acquired thermotolerance in *Arabidopsis thaliana*. Plant J **38:** 432–447

Cook D, Fowler S, Fiehn O, Thomashow MF (2004) A prominent role for the CBF cold responsive pathway in configuring the low-temperature metabolome of *Arabidopsis*. Proc Natl Acad Sci USA **101:** 15243–15248

Dat JF, Foyer CH, Scott IM (1998) Changes in salicylic acid and antioxidants during induced thermotolerance in mustard seedlings. Plant Physiol **118:** 1455–1461

Dixon RA (2001) Natural products and plant disease resistance. Nature **411:** 843–847

Fan TW-M, Colmer TD, Lane AN, Higashi RM (1993) Determination of metabolites by ^{1}H NMR and GC: analysis for organic osmolytes in crude tissue extracts. Anal Biochem **214:** 260–271

Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L (2000) Metabolite profiling for plant functional genomics. Nat Biotechnol **18:** 1157–1161

Fowler S, Thomashow MF (2002) Arabidopsis transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway. Plant Cell **14:** 1675–1690

Garcia AB, Engler J, Iyer S, Gerats T, Van Montagu M, Caplan AB (1997) Effects of osmoprotectants upon NaCl stress in Rice. Plant Physiol **115:** 159–169

Gates DM (1968) Transpiration and leaf temperature. Annu Rev Plant Physiol **19:** 211–238

Gilmour SJ, Sebolt AM, Salazar MP, Everard JD, Thomashow MF (2000) Overexpression of the Arabidopsis CBF3 transcriptional activator mimics multiple biochemical changes associated with cold acclimation. Plant Physiol **124:** 1854–1865

Guy CL (1999) Molecular responses of plants to cold shock and cold acclimation. J Mol Microbiol Biotechnol **12:** 231–242

Guy CL, Niemi KJ, Brambl R (1985) Altered gene expression during cold acclimation of spinach. Proc Natl Acad Sci USA **82:** 3673–3677

Guy CL, Huber JLA, Huber SC (1992) Sucrose phosphate synthase and sucrose accumulation at low temperature. Plant Physiol **100:** 502–508

Hallberg RL, Kraus KW, Hallberg EM (1985) Induction of acquired thermotolerance in *Tetrahymena thermophila*: effects of protein synthesis inhibitors. Mol Cell Biol **8:** 2061–2069

Handa S, Bressan RA, Handa AK, Carpita NC (1983) Solutes contributing to osmotic adjustment in cultured plant cells adapted to water stress. Plant Physiol **73:** 834–843

Heil M, Bostock RM (2002) Induced systemic resistance (ISR) against pathogens in the context of induced plant defenses. Ann Bot (Lond) **89:** 503–512

Klee H (2003) Hormones are in the air. Proc Natl Acad Sci USA **100:** 10144–10145

Koch KE (1996) Carbohydrate-modulated gene expression in plants. Annu Rev Plant Physiol Plant Mol Biol **47:** 509–540

Kreps JA, Wu Y, Chang H-S, Zhu T, Wang X, Harper JF (2002) Transcriptome changes for Arabidopsis in response to salt, osmotic and cold stress. Plant Physiol **130:** 2129–2141

Levitt J (1972) Responses of plants to environmental stresses. Academic Press, New York

Lopez-Delgado H, Dat J, Foyer C, Scott I (1998) Induction of thermotolerance in potato microplants by acetylsalicylic acid and H_2O_2. J Exp Bot **49:** 713–720

Lurie S, Klein JD (1991) Acquisition of low-temperature tolerance in tomatoes by exposure to high temperature stress. J Am Soc Hortic Sci **116:** 1007–1012

Malamy J, Carr JP, Klessig DF, Raskin I (1990) Salicylic acid: a likely endogenous signal in the resistance response of tobacco to viral infection. Science **250:** 1002–1004

Mayer RR, Cherry JH, Rhodes D (1990) Effects of heat shock on amino acid metabolism of cowpea cells. Plant Physiol **94:** 796–810

McCollum TG, Doostdar H, Mayer RT, McDonald RE (1995) Immersion of cucumber fruit in heated water alters chilling-induced physiological changes. Postharvest Biol Technol **6:** 55–64

McKown R, Kuroki G, Warren G (1996) Cold responses of *Arabidopsis* mutants impaired in freezing tolerance. J Exp Bot **47:** 1919–1925

Metraux JP, Signer H, Ryals J, Ward E, Wyss-Benz M, Gaudin J, Raschdorf K, Schmid E, Blum W, Inverardi B (1990) Increase in salicylic acid at the onset of systemic acquired resistance in cucumber. Science **250:** 1004–1006

Mittler R (2002) Oxidative stress, antioxidants and stress tolerance. Trends Plant Sci **9:** 405–410

Moore B, Zhou L, Rolland F, Hall Q, Cheng W-H, Liu Y-X, Hwang I, Jones T, Sheen J (2003) Role of the Arabidopsis glucose sensor HXK1 in nutrient, light, and hormonal signaling. Science **300:** 332–336

Munne-Bosch S, Penuelas J (2003) Photo- and antioxidative protection, and a role for salicylic acid during drought and recovery in field-grown *Phillyrea angustifolia* plants. Planta **217:** 758–766

Rizhsky L, Liang H, Shuman J, Shulaev V, Davletova S, Mittler R (2004) When defense pathways collide. The response of Arabidopsis to a combination of drought and heat stress. Plant Physiol **134:** 1683–1696

Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L (2000) Technical advance: simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. Plant J **23:** 131–142

Roitsch T (1999) Source-sink regulation by sugar and stress. Curr Opin Plant Biol **2:** 198–206

Rolland F, Moore B, Sheen J (2002) Sugar sensing and signaling in plants. Plant Cell **14:** 185–205

Saltveit ME (2002) Heat shocks increase the chilling tolerance of rice (*Oryza sativa*) seedling radicles. J Agric Food Chem **50:** 3232–3235

Saltveit ME, Hepler PK (2004) Effect of heat shock on the chilling sensitivity of trichomes and petioles of African violet (*Saintpaulia ionantha*). Physiol Plant **121:** 35–43

Saltveit ME, Peiser G, Rab A (2004) Effect of acetaldehyde, arsenite, ethanol, and heat shock on protein synthesis and chilling sensitivity of cucumber radicles. Physiol Plant **120:** 556–562

Schmelz EA, Engelberth J, Alborn HT, O'Donnell P, Sammons M, Toshima H, Tumlinson JH III (2003) Simultaneous analysis of phytohormones, phytotoxins, and volatile organic compounds in plants. Proc Natl Acad Sci USA **100:** 10552–10557

Seki M, Narusaka M, Abe H, Kasuga M, Yamaguchi-Shinozaki K, Carninci P, Hayashizaki Y, Shinozaki K (2001) Monitoring the expression pattern of 1300 Arabidopsis genes under drought and cold stresses by using a full-length cDNA microarray. Plant Cell **13:** 61–72

Senaratna T, Touchell D, Bunn E, Dixon K (2000) Acetyl salicylic acid (aspirin) and salicylic acid induce multiple stress tolerance in bean and tomato plants. Plant Growth Regul **30:** 157–161

Shahjee HM, Banerjee K, Ahmad F (2002) Comparative analysis of naturally occurring L-amino acid osmolytes and their D-isomers on protection of *Escherichia coli* against environmental stresses. J Biosci **27:** 515–520

Shinozaki K, Dennis ES (2003) Cell signalling and gene regulation: global analyses of signal transduction and gene expression profiles. Curr Opin Plant Biol **6:** 405–409

Smeekens S (2000) Sugar-induced signal transduction in plants. Annu Rev Plant Physiol Plant Mol Biol **51:** 49–81

Srivastava KK, Sinha RK, Pandey PK, Prasad M (1980) Variations in amino acids and sugars in different tissues of broad bean (*Vicia faba* L.) during pathogenesis of *Uromyces fabae* (Pers.) de Bary. Zentralbl Bakteriol Naturwiss **135:** 344–350

Sung DY, Guy CL (2003) Physiological and molecular assessment of altered expression of Hsc70-1 in Arabidopsis. Evidence for pleiotropic consequences. Plant Physiol **132:** 979–987

Taji T, Ohsumi C, Iuchi S, Seki M, Kasuga M, Kobayashi M, Yamaguchi-Shinozaki K, Shinozaki K (2002) Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance in *Arabidopsis thaliana*. Plant J **29:** 417–426

Thomashow MF (1999) Plant cold acclimation: freezing tolerance genes and regulatory mechanisms. Annu Rev Plant Physiol Plant Mol Biol **50:** 571–599

Uemura M, Warren G, Steponkus PL (2003) Freezing sensitivity in the sfr4 mutant of Arabidopsis is due to low sugar content and is manifested by loss of osmotic responsiveness. Plant Physiol **131:** 1800–1807

Vetter J (2000) Plant cyanogenic glycosides. Toxicon **38:** 11–36

Wagner C, Sefkow M, Kopka J (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. Phytochemistry **62:** 887–900

Xin Z, Browse J (1998) Eskimo1 mutants of Arabidopsis are constitutively freezing-tolerant. Proc Natl Acad Sci USA **95:** 7799–7804

Yancey PH, Clark ME, Hand SC, Bowlus RD, Somero GN (1982) Living with water stress: evolution of osmolyte systems. Science **217:** 1214–1222

Yelenosky G, Guy CL (1977) Carbohydrate accumulation in leaves and stems of "Valencia" orange at progressively colder temperatures. Bot Gaz **138:** 13–17

The Plant Journal (2007) **50**, 967–981 doi: 10.1111/j.1365-313X.2007.03100.x

Transcript and metabolite profiling during cold acclimation of Arabidopsis reveals an intricate relationship of cold-regulated gene expression with modifications in metabolite content

Fatma Kaplan[1], Joachim Kopka[2], Dong Yul Sung[1], Wei Zhao[3], Mick Popp[4], Ron Porat[5] and Charles L. Guy[1,*]

[1]*Plant Molecular and Cellular Biology Program, Department of Environmental Horticulture, University of Florida, Gainesville, FL 32611, USA,*

[2]*Max Planck Institute of Molecular Plant Physiology, 14476 Golm, Germany,*

[3]*Department of Statistics, University of Florida, Gainesville, FL 32611, USA,*

[4]*Interdisciplinary Center for Biotechnology Research, Box 100156, Gainesville, FL 32610, USA, and*

[5]*Department of Postharvest Science, ARO, the Volcani Center, Bet Dagan, Israel*

Received 19 January 2007; accepted 15 February 2007

*For correspondence (fax +1 352 392 1413; e-mail clguy@ufl.edu).

Summary

Exposure of Arabidopsis to low temperatures results in cold acclimation where freezing tolerance is enhanced. To achieve a wider view of the role of transcriptome to biochemical changes that occur during cold acclimation, analyses of concurrent transcript and metabolite changes during cold acclimation was performed revealing the dynamics of selected gene–metabolite relationships. Exposure to low temperature resulted in broad transcriptional and metabolite responses. Principal component analysis revealed sequentially progressive, global changes in both gene expression and metabolite profiles during cold acclimation. Changes in transcript abundance for many metabolic processes, including protein amino acid biosynthetic pathways and soluble carbohydrates, during cold acclimation were observed. For some metabolic processes, changes in transcript abundance temporally correlated with changes in metabolite levels. For other metabolic processes, changes in transcript levels were not correlated with changes in metabolite levels. The present findings demonstrate that regulatory processes independent of transcript abundance represent a key part of the metabolic adjustments that occur during cold acclimation.

Keywords: post-transcriptional regulation, low temperature, metabolism, sugars, proline, GABA.

Introduction

The ability of plants to survive and recover from unfavorable or stressful conditions is a function of basal and acquired tolerance mechanisms (Levitt, 1972). With Arabidopsis, it has been shown that exposure to low non-freezing temperatures enhances freezing tolerance, a phenomenon known as cold acclimation (CA) (Thomashow, 1999). As such, this acquired freezing tolerance is a complex biological process involving the activation of various molecular, biochemical and physiological changes, including changes in membrane structure and function, tissue water content, and global changes in gene expression, protein, lipid, and primary and secondary metabolite composition (Gilmour *et al.*, 2000; Guy, 1990; Guy *et al.*, 1992; Levitt, 1972; Renaut *et al.*, 2006; Thomashow, 1999; Wang *et al.*, 2006).

Recent advances in genome sequencing and global gene expression analysis have helped to further establish the multigenic nature of plant responses to environmental stresses, and reveal hundreds of genes whose expression patterns have been linked to CA processes (Fowler and Thomashow, 2002; Kreps *et al.*, 2002; Seki *et al.*, 2002; Shinozaki and Dennis, 2003; Vogel *et al.*, 2005). Similarly, non-targeted metabolite profiling analysis allows the simultaneous monitoring of precursors, intermediates and products of metabolic pathways that can be linked with CA. Indeed, metabolite profiling using various forms of mass spectrometry has revealed changes in the steady-state pools of more than 300 polar metabolites in response to cold shock, CA (Cook *et al.*, 2004; Kaplan *et al.*, 2004) and

low-temperature development (Gray and Heath, 2005). In an examination of natural variation in freezing tolerance, using a collection of Arabidopsis accessions representing ecotypes originating from high latitude cold climates to low latitude warm climates, based on metabolite and transcript profiling of non-acclimated and acclimated plants, it was concluded that global changes in metabolite profiles do not correlate with the ability to cold acclimate, whereas global changes in transcriptome profiles do not correlate with the ability to cold acclimate (Hannah *et al.*, 2006).

A concomitant linkage between changes in the transcriptome and the metabolome during CA is expected, but has received little attention. One study by Cook *et al.* (2004) has shown that as much as 79% of the metabolite changes elicited during CA were also in common in non-acclimated plants in response to overexpression of the transcription factor (C-repeat/dehydration responsive element-binding factor 3, CBF3) that acts to regulate a major cold stress regulon linked with enhanced freezing tolerance (Vogel *et al.*, 2005). In the study by Hannah *et al.* (2006), it appears that enhanced freezing tolerance is linked with the down-regulation of genes for photosynthesis, but with the induction of genes involved with flavonoid metabolism. However, the involvement of post-transcriptional regulation of metabolism, although well-known in the biological world, has not yet been well established as playing a role in the reconfiguration of the metabolome during CA. Given the multiplicity of the regulatory mechanisms expected to control metabolism during CA, it is clear that by combining gene expression and metabolite profiling analyses it should be possible to identify which metabolite responses are influenced by changes in transcript abundance, those responses controlled by mechanisms independent of transcript abundance and those that may have elements of both.

Combining genome-wide expression and metabolite profiling to examine the linkage between changes in gene expression and changes in metabolite level offers a systems approach (Bohnert *et al.*, 2006; Cramer *et al.*, 2007; Kirschner, 2005; Oksman-Caldentey and Saito, 2005; Oksman-Caldentey *et al.*, 2004) to better understand the complex processes occurring during the onset of CA. Using such an integrated approach, Hirai *et al.* (2004, 2005) was able to identify new genes involved in glucosinolate metabolism in response to sulfur depletion, and Tohge *et al.* (2005) identified novel flavonoid biosynthesis pathways in transgenic Arabidopsis plants overexpressing a MYB transcription factor.

In the present study, we conducted analyses of datasets from transcript and metabolite profiling. Experiments were originally conducted where the same tissue samples for gene expression analysis and metabolite profiling were prepared. The metabolite profiling dataset used in the present work has been previously published (Kaplan *et al.*, 2004). By integration of gene expression and metabolite profiling, a more complete understanding of the diverse nature of the dynamics of transcript abundance–metabolite linkages of selected biochemical pathways activated during the onset of CA is revealed. We show here that metabolic reconfiguration in response to CA emanates from both changes in transcript abundance and regulatory processes independent of transcript abundance.

Results

Temporal transcript and metabolite steady-state levels during cold acclimation

DNA microarray analysis resulted in the identification of 8171 probe sets that showed statistically significant signal intensity differences affected by the CA treatment (Table S1). To detect changes in transcript and metabolite content during the onset of CA, plants were kept at 4°C for up to 96 h. The metabolite data used in the present study were obtained from our previously published work (Kaplan *et al.*, 2004). All metabolite data used in the current analyses was obtained from the same plant material used for conducting gene expression analyses, creating experimentally matched datasets to make this intregative analysis possible. Sustained and transient responses were classified into three major categories: early (changes occurred over 1–4 h), intermediate (changes occurred over 12–24 h) and late (changes occurred over 48–96 h). A portion of the observed changes in gene expression included transcripts for genes encoding enzymes involved in many cellular biosynthetic pathways, such as the induction of callose, fermentation, phospholipid, starch, sugar, flavonoid, protein amino acids, γ-aminobutyric acid (GABA) and terpenoid biosynthesis, and the repression of photorespiration, folic acid, betaine, sulfate assimilation, ethylene, fatty acid, gluconeogenesis, amino acids, brassinosteroids and chlorophyll biosynthesis (the list of genes in this analysis belonging to the different metabolic pathways is presented in Table S2). Similarly, parallel metabolite profiling by GC-MS on the same plants used for gene expression analyses revealed that the signal intensities of 311 low-molecular mass polar compounds were altered in response to CA, as previously reported (Kaplan *et al.*, 2004). Overall, the relative steady-state pool sizes of amino acids, Tricarboxylic acid (TCA) cycle intermediates, and many metabolites of carbohydrate metabolism and phenylpropanoid pathway intermediates were affected by CA (Table S3).

Analyses of the statistically significant variations for transcripts and metabolites in the temporal responses of biosynthetic pathways monitored during CA were conducted, and revealed four types of response relationships: 1, increases in transcript signal levels for enzymes of some pathways (raffinose biosynthesis and GABA shunt) that preceded increases in metabolite signal levels in a fashion

consistent with transcriptional activation; 2, for many biosynthetic pathways, including Gly, Ala, Thr, Leu, sucrose synthesis and the TCA cycle, increases in metabolite signal levels preceded increases in transcript abundance (Table 1); 3, for other pathways such as Lys, Met, Trp, Tyr, Arg, Cys, polyamine and phenylpropanoid biosynthetic pathways, transcript signal levels of many genes decreased, whereas their corresponding metabolite signal intensities increased (Table 1); and 4, transcripts for enzymes of the biosynthetic pathways for Ile, Val and Glu did not show detectable changes, but the corresponding metabolite content showed considerable increases. These four types of temporal response relationships for transcript and metabolite data demonstrate a variety of gene–metabolite linkages in plant responses to CA that are transcript abundance dependent and independent.

Principal component analysis

Principal component analysis (PCA) is used to reduce multivariate data complexity as a method of identifying patterns, and expressing data, in ways that highlight similarities and differences. A common use of PCA has been to demonstrate that composite metabolite profiles can be characteristic of the metabolic status of a plant or tissue: i.e. reveal a characteristic and perhaps diagnostic 'metabolic phenotype' (Fiehn *et al.*, 2000). A similar concept can also be applied to the gene expression status (e.g. Scholz *et al.*, 2005). We used PCA to evaluate simultaneous changes in global gene expression patterns (this study) and metabolite signal intensity profile data that were obtained from a previously published work (Kaplan *et al.*, 2004). The results show that during the 4-day time course of CA, neither the transcript nor the metabolite patterns were static, but followed a progressive sequence, indicative of continuous change (Figure 1) in both the transcriptome expression profile and the polar metabolite profile.

Amino acid metabolism

The signal levels of the majority of protein amino acids increased during the time-course of CA, whereas transcript

Table 1 Summary of general steady-state transcript[a] and metabolite level changes during the course of 96 h cold acclimation

Transcript level	Metabolite level	Pathway
Increase early sustained	Increase early sustained	SERINE BIOSYNTHESIS
Increase early sustained	Increase early sustained	SUCROSE BIOSYNTHESIS
Increase early sustained	Increase early sustained	STARCH DEGRADATION HYDROLYTIC
Increase intermediate	Increase early sustained	GLYCINE BIOSYNTHESIS
Increase intermediate	Increase early sustained	GLYCOLYSIS
Increase intermediate sustained	Increase early sustained	ALANINE BIOSYNTHESIS
Increase intermediate sustained	Increase early sustained	TCA CYCLE
Increase intermediate sustained	Increase early sustained	THREONINE BIOSYNTHESIS
Increase intermediate transient	Increase early sustained[b]	PROLINE BIOSYNTHESIS
Increase late	NA[c]	PROLINE DEGRADATION
Increase intermediate transient (amino transferase)	Increase early sustained	LEUCINE BIOSYNTHESIS
No significant change	Increase early sustained	ISOLEUCINE BIOSYNTHESIS
No significant change	Increase early sustained	VALINE BIOSYNTHESIS
Decrease intermediate transient	NA	ISOLEUCINE VALINE DEGRADATION
Decrease intermediate	Increase early sustained	LYSINE BIOSYNTHESIS
Decrease intermediate	Increase early sustained	METHIONINE BIOSYNTHESIS
Decrease intermediate	Increase early sustained	POLYAMINE BIOSYNTHESIS
Decrease intermediate and late	Increase early sustained	TRYPTOPHAN BIOSYNTHESIS
Decrease intermediate and late	Increase early sustained	TYROSINE BIOSYNTHESIS
Decrease intermediate transient	Increase early sustained	ARGININE BIOSYNTHESIS
Decrease late	Increase early sustained	CYSTEINE BIOSYNTHESIS
Increase intermediate transient	Increase early sustained U shape	TREHALOSE BIOSYNTHESIS
Increase intermediate transient	Increase intermediate sustained	RAFFINOSE BIOSYNTHESIS
No significant change	Increase intermediate sustained	ASPARAGINE BIOSYNTHESIS
No significant change	Increase intermediate sustained	GLUTAMATE BIOSYNTHESIS
Decrease early sustained	NA	GLUTAMATE DEGRADATION
Decrease early and intermediate	Increase intermediate sustained	PHENYLPROPANOID PATHWAY
Increase intermediate sustained	Increase intermediate transient 24 h	GABA SHUNT
Decrease intermediate	Decrease Intermediate Transient	ASPARTATE BIOSYNTHESIS

Early increase/decrease; the first statistically significant response occurred either at 1 or 4 h. Intermediate increase/decrease; the first statistically significant response occurred at 12 or 24 h. Late increase/decrease; the first statistically significant response occurred at 48 or 96 h.
[a]A list of genes exhibiting changes in transcript abundance associated with each pathway can be found Table S2. [b]Diurnal regulation; [c]not available

 The Plant Journal, (2007), **50**, 967–981

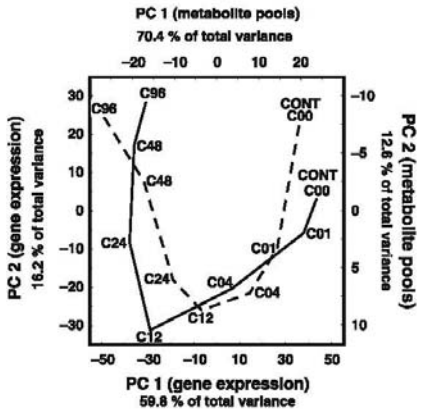


Figure 1. A parallel principal component analysis (PCA) of transcript and metabolite profiles during the time course of cold acclimation.
Each point represents means of three experiments for gene expression and two experiments for metabolite profiling. Transcript PCA is indicated by the solid line and metabolite PCA is indicated by the dashed line. The various time points (in h) are indicated within the graph. The metabolite data used for this analysis was obtained from a previously published dataset (Kaplan *et al.*, 2004).

signal levels involved in amino acid biosynthesis increased, remained unchanged or decreased (Table 1). An illustration of the general organization of amino acid biosynthetic pathways is shown in Figure 2. Apparent coordinate increases in the pool sizes of amino acids derived from 3-phosphoglycerate (Ser, *O*-acetylserine, Cys), phosphoenolpyruvate (Phe, Trp), pyruvate (Leu, Val), oxaloacetate (Asn, Lys, Met, Thr, Ile) and α-ketoglutarate (Glu, Pro, Gln, Arg) were observed during CA (Figure 2). Notable exceptions to the coordinate increases were pathway intermediates shikimate and Asp, which remained unchanged at time-zero levels for the duration of CA. Clearly the considerable increase in the aromatic amino acids was not dependent on increased steady-state levels of shikimate.

Among the protein amino acids, proline has attracted much attention because it robustly accumulates under stress conditions including cold stress (Gilmour *et al.*, 2000; Verbruggen *et al.*, 1993; Wanner and Junttila, 1999; Xin and Browse, 1998; Yoshiba *et al.*, 1995). There are two biosynthetic pathways that lead to Pro: one from glutamate (Figure 3) and the other from ornithine. In the glutamate pathway, the major enzyme providing the rate-controlling step in Pro biosynthesis in plants is Δ^1-pyrroline-5-carboxylate synthetase (*P5CS*), which catalyzes the first two steps (Figure 3a) from precursor Glu to Δ^1-pyrroline-5-carboxylate (P5C). This enzyme is subject to feedback regulation by Pro and by light-mediated changes in gene expression (Hayashi *et al.*, 1996; Hu *et al.*, 1992). Arabidopsis contains two *P5CS* genes (Strizhov *et al.*, 1997), both of which are represented on the ATH1 array as a single probe set. Transcript and metabolite profiling here revealed that the steady-state mRNA signal levels for *P5CS* markedly increased after 12 h of exposure to 4°C, and then declined to near pre-stress levels after 96 h of cold (Figure 3b). Transcript signal levels for Δ^1-pyrroline-5-carboxylate reductase (*P5CR*) (At5 g14800) were unchanged during the first 24 h of exposure to low temperature, and then modestly increased to 1.3- and 1.7-fold after 48 and 96 h (not shown). In contrast with the observed collective transcript signal levels, a statistically significant increase of more than twofold in Pro signal intensity occurred after 4 h of exposure to 4°C, and this was followed by a continuous and dramatic increase to 130-fold of the control signal level after 96 h (Figure 3b). The very early increase in the Pro signal appears to result from diurnal variation, which was unaltered by low temperature as untreated control plants taken at the same time point (4-h control) exhibited a similar signal level for Pro (data not shown).

During Pro degradation, proline dehydrogenase (PDH) catalyzes a darkness-regulated conversion of Pro to P5C (Hayashi *et al.*, 1996), and Δ^1-pyrroline-5-carboxylate dehydrogenase (P5CDH) produces Glu from P5C (Figure 3). Following an initial decrease, transcript signal levels for *PDH* (At3 g30775) continuously increased during the light period (Figure 3) by 4.5-fold after 96 h of CA, whereas the transcript signal levels for *P5CDH* (At5g62530) were unchanged by CA after 48 and 96 h, when the levels were slightly increased. Taken together, it is clear that during CA, modulation of transcript abundance is not sufficient to explain the dynamics in Pro levels. Modulation of enzymatic activity must also be a vital regulatory process that integrates with changes in gene expression.

GABA shunt

GABA is a four-carbon amine-containing metabolite with known cryoprotective properties that has been suggested to play a role in stress-signaling responses (Bouche and Fromm, 2004). GABA is synthesized by the irreversible decarboxylation of the amino acid Glu by a cytosolic glutamate decarboxylase (GAD) (Shelp *et al.*, 1999; Figure 4a). Its degradation is further coordinated by the activity of two mitochondrial enzymes: GABA transaminase (GABA-T), which catalyzes a reversible reaction between GABA and succinic semialdehyde (SSA), and succinic semialdehyde dehydrogenase (SSADH), which irreversibly oxidizes SSA to succinate (Bouche *et al.*, 2003) (Figure 4a). Increases in transcript signal levels for two *GAD* genes were apparent by 12 h of CA, and preceded the increase in GABA signal intensity that peaked by 24 h, thus demonstrating a characteristic transcript abundance-regulated response (Figure 4b).

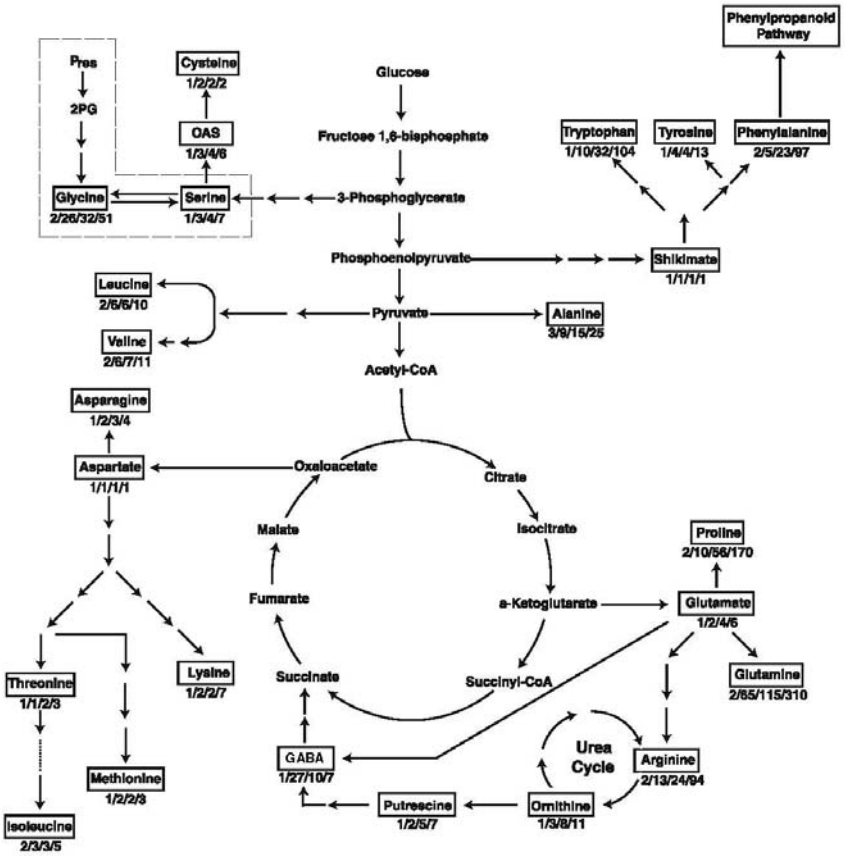


Figure 2. Change in amino acid steady-state signal levels in the context of amino acid biosynthetic families.
The values below each metabolite box indicate the -fold change in signal intensity relative to time zero for 1, 24, 48 and 96 h of cold acclimation. The broken line indicates the photorespiratory (P_{res}) pathway for Gly and Ser biosynthesis. GABA, γ-aminobutyric acid; OAS, *O*-acetylserine; 2PG, 2-phosphoglycollate. This figure was modified with permission from Figure 8.2 in Coruzzi and Last (2000) to illustrate signal intensity dynamics during cold acclimation.

Regarding GABA catabolism, we found that *GABA-T* transcript signal levels were unchanged as a result of CA, but that *SSADH* transcript signal levels peaked after 24 h and thus preceded the following decrease in GABA content observed by 48 h (Figure 4b). Succinic semialdehyde can also be reduced to 4-hydroxybutarate (GHB) by the reversible action of cytosolic succinic semialdehyde reductase (SSR). GHB was detected in cold acclimated plants, consistent with the rise and fall of GABA, but the transcript signal levels for *SSR* were slightly downregulated overall during CA (not shown). Therefore, transcript and metabolite profiling suggest that the transient increase in GABA content at 24 h was tightly linked to the coordinate increase of *GAD* and *SSADH* transcript levels during CA.

Sucrose metabolism

Precursors for sucrose synthesis derive from the hexose phosphate pool, consisting of Glc-1-P, Glc-6-P and Fru-6-P, and the sugar nucleotide UDPGlc (Figure 5a). The three hexose phosphates are maintained in a relative steady-state balance by phosphoglucomutase, and by Glc-6-P isomerase (Dennis and Blakeley, 2000). Upon transfer to low temperature, Glc-6-P and Fru-6-P signal levels rapidly increased hyperbolically until 12 h, and then remained elevated during

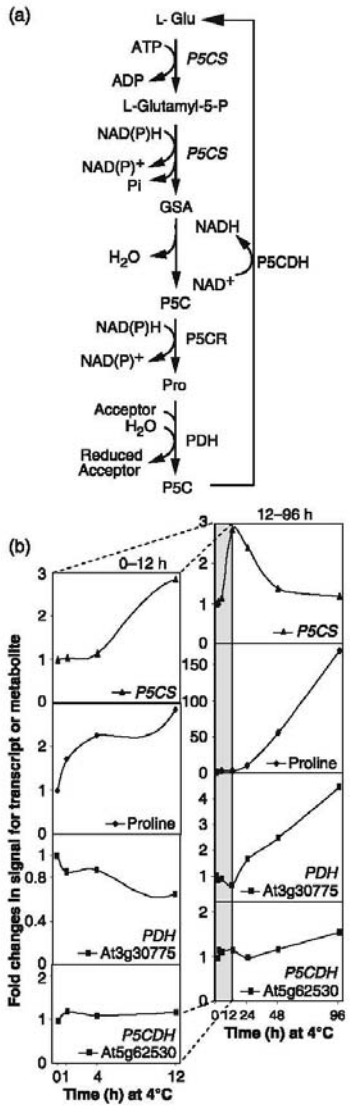


Figure 3. Integration of transcript and metabolite profiling for Pro metabolism during the cold acclimation (CA) time course.
(a) Overview of proline biosynthesis and degradation.
(b) Metabolite and transcript signal levels. The figures on the left are enlargements of the changes in transcript and metabolite signal intensities accruing during the early stages of CA (0–12 h). GSA, glutamic γ-semialdehyde; *P5C*, Δ^1-pyrroline-5-carboxylate; P5CDH, Δ^1-pyrroline-5-carboxylate dehydrogenase; P5CR, Δ^1-pyrroline-5-carboxylate reductase; P5CS, Δ^1-pyrroline-5-carboxylate synthetase; PDH, proline dehydrogenase. Metabolite data were obtained as described in the text.

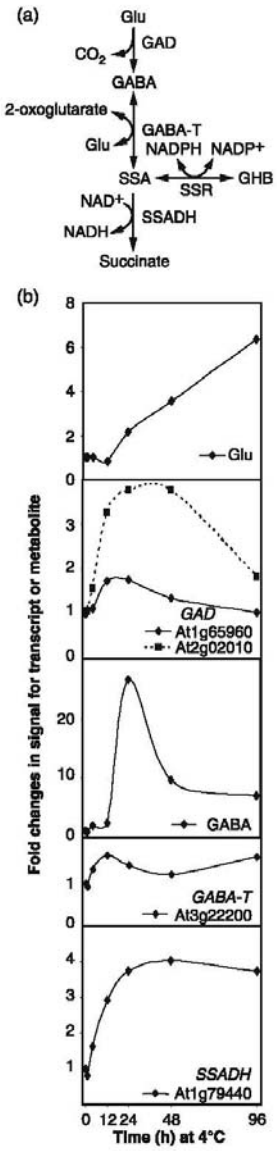


Figure 4. Integration of transcript and metabolite profiling for GABA metabolism during the cold acclimation time course.
(a) Overview of GABA shunt.
(b) Metabolite and transcript signal levels. GABA, γ-aminobutyric acid; GABA-T, GABA transaminase; GAD, glutamate decarboxylase; GHB, 4-hydroxybutyric acid; SSA, succinic semialdehyde; SSADH, SSA dehydrogenase; SSR, succinic semialdehyde reductase. Metabolite data were obtained as described in the text.

 The Plant Journal, (2007), **50**, 967–981

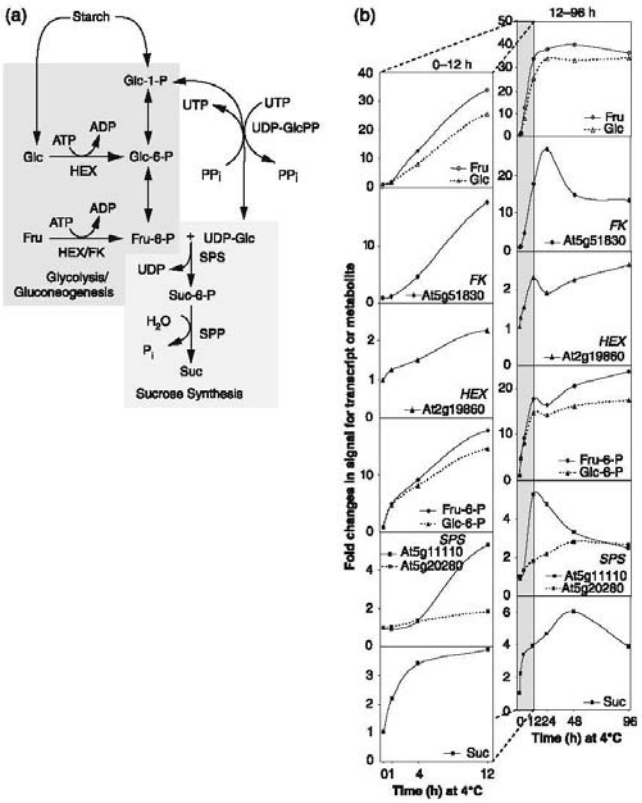


Figure 5. Integration of transcript and metabolite profiling of hexose phosphates and sucrose synthesis during the cold acclimation (CA) time course.
(a) Brief overview of hexose-phosphate and sucrose synthesis.
(b) Metabolite and transcript signal levels. The figures on the left are enlargements of the changes for transcript and metabolite levels accrued during the early stages of CA (0–12 h). Fru, fructose; FK, fructokinase; Glc, glucose; HEX, hexokinase; SPP, sucrose phosphate phosphatase; SPS, sucrose phosphate synthase; SS, sucrose synthase; Suc, sucrose; UDP-GlcPP; UDP-glucose pyrophosphorylase. Metabolite data were obtained as described in the text.

the duration of CA (Figure 5b). Consistent with the mass action functions of phosphoglucomutase and Glc-6-P isomerase, their respective transcript signal levels did not exhibit any noteworthy changes during 96 h of CA (data not shown). These enzymes function as equilibrators, and thus account for the tight parallelism of the Glc-6-P and Fru-6-P levels throughout the low-temperature treatment. Upon transfer to low temperature, Fru-6-P signal levels remained slightly greater than that of Glc-6-P. However, the activity of Glc-6-P isomerase should favor greater levels of Glc-6-P over that of Fru-6-P.

Precursors of the hexose phosphate pool may come from a gluconeogenesis process by phosphorylating free hexoses produced from starch and/or sucrose degradation (Figure 5a). The enzymes hexokinase (HEX) and fructokinase (FK) catalyze the phosphorylation of free Glc and Fru, respectively (Figure 5a). A moderate increase in *HEX* transcript signal levels, which paralleled hexose phosphate increases during the 96-h period of CA, was detected (Figure 5b). In contrast, the transcript signal levels for *FK* dramatically increased by 4 h of cold shock, and continued to increase for up to 24 h, before decreasing thereafter (Figure 5b). The kinetics of the robust induction of *FK* in relation to the rise in Fru-6-P levels suggests that, in addition to glycolytic and/or gluconeogenic processes, FK may provide another biosynthetic source that contributes to the increases in the steady-state pool of Fru-6-P.

Sucrose phosphate synthase (SPS) catalyzes sucrose synthesis by converting UDP-Glc and Fru-6 to produce Suc-6-P, and this reaction is followed by the activity of sucrose phosphate phosphatase (SPP), which removes the phosphate group from Suc-6-P to produce Suc and inorganic phosphate (Figure 5a). We found that *SPS* transcript signal levels increased and peaked at 12 h of CA, and afterwards declined to levels still well above those present prior to low-temperature exposure (Figure 5b). Transcript signal levels for *SPP* were unchanged during CA (data not shown). Sucrose signal levels increased immediately within 1 h of exposure to cold shock, and continued to increase until 48 h, and then afterwards decreased at 96 h

 The Plant Journal, (2007), **50**, 967–981

(Figure 5b). To rule out the possibility that the threefold increase in sucrose signal level after 4 h of CA (6 h into the light period) was a photosynthetically driven diurnal change, sucrose signal levels were measured in control plants at the same time point, but without exposure to cold temperatures. We found no change in sucrose content in the control plants (data not shown). These results indicate that the early increase in sucrose signal levels in response to CA preceded the observed increase in *SPS* and *SPP* transcript signal levels. Therefore, the initial increase in sucrose was not dependent on increased transcript abundance.

To examine whether the increase in sucrose content during CA may be a result of a decrease in its utilization and degradation, the transcript signal levels for invertases and sucrose synthases (SS) were evaluated. It was found that the transcript signal levels encoding several invertases increased, whereas those of other invertases decreased. Similarly, the transcript signal levels of most *SS* genes were unaffected by CA, and only the transcript signal of a single gene showed a transient increase (data not shown).

Raffinose metabolism

Raffinose is a trisaccharide composed of galactose, glucose and fructose units, and functions as a major compatible solute in abiotic stress responses of plants (Taji *et al.*, 2002). The immediate precursors of raffinose are galactinol and sucrose, and its biosynthesis rate depends on their availability (Figure 6a). Galactinol is synthesized from *myo*-inositol and UDP-galactose by galactinol synthase, and then galactinol reacts with sucrose to form raffinose and *myo*-inositol catalyzed by raffinose synthase (Taji *et al.*, 2002; Figure 6a). Indeed, as seen in Figures 5(b) and 6(b), increases in sucrose and galactinol levels clearly precede the increase in raffinose content, as might be expected from substrate–product relationships. We found that transcript signal levels for genes of galactinol synthase and raffinose synthase increased during CA, and were coordinately regulated. Transcript signal levels for both enzymes increased after 4 h of cold shock, synchronously peak at 12 h, and then gradually declined to low steady-state levels at 96 h (Figure 6b). Perhaps, as a consequence, or as an advantage of the low-temperature influence on metabolism, it can be seen that the increases in the transcript signal levels for genes encoding both galactinol synthase and raffinose synthase also clearly preceded the later increases in galactinol and raffinose signals, which became evident only at 12 and 24 h, respectively, exactly as would be predicted from a transcript abundance-regulated process (Figure 6b).

Raffinose degradation results in either the production of fructose and mellibiose or sucrose and galactose (Figure 6c). Mellibiose signal intensity did not significantly change until 96 h (1.3-fold) of CA, and its level remained the same even after 24 h of deacclimation (1.4-fold), when raffinose signals returned to near-control levels (2.6-fold) (Kaplan *et al.*, 2004). Similarly, the levels of another alternative degradation product, sucrose, also decreased after 24 h of deacclimation when raffinose was being rapidly metabolized (Kaplan *et al.*, 2004).

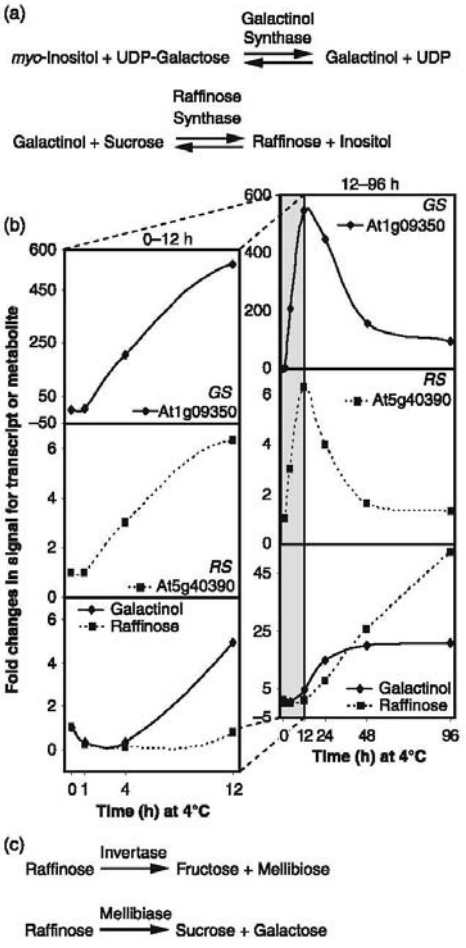


Figure 6. Integration of transcript and metabolite profiling of raffinose biosynthetic components during the cold acclimation (CA) time course.
(a) Raffinose biosynthesis.
(b) Time course of metabolite and gene expression. The figures on the left are enlargements of the changes in transcript and metabolite levels accrued during the early stages of CA (0–12 h).
(c) Raffinose degradation products. GS, galactinol synthase; RS, raffinose synthase. Metabolite data were obtained as described in the text.

Time course and dynamics of carbohydrate responses

Many studies have shown that the content of soluble sugars in Arabidopsis leaves increase dramatically during CA (Cook *et al.*, 2004; Gilmour *et al.*, 2000; Kaplan *et al.*, 2004; Rohde *et al.*, 2004; Wanner and Junttila, 1999). In the current study, based on metabolite profiling analysis during the CA time course, the system dynamics and sequence of events leading to enhanced carbohydrate levels were examined. Figure 7 shows the time course of signal responses for sugars during CA, and through interpolation designates the relative timing of the rise or decline in signal value for each metabolite by calculating the time point where it reached 50% in its maximum or minimum value ($T_{0.5[MAX]}$). It can be seen that the first carbohydrates to show increases during CA were maltose and maltotriose, which are the direct breakdown products of starch degradation by β-amylase activity (Figure 7b). The maltotriose signal increase followed that of maltose and, therefore, it may either be a derivative or a consequence of maltose disproportionation to maltotriose and glucose. The signal levels of the disaccharides isomaltose and maltitol show little change and do not appear to be major accumulated products of starch degradation. Their signal levels rise later than maltose and maltotriose, and seem to be linked with high signal levels of maltose and maltotriose (Figure 7a). These increases for starch breakdown products were synchronously followed by the buildup of the hexose phosphates Glc-6-P, Fru-6-P, Gal-6-P and Man-6-P (Figure 7b). The signal levels of Gal-6-P and Man-6-P rapidly rise, exhibiting a striking parallel with each other followed by a two-cycle oscillating return to pre-stress levels (Figure 7b). The responses of Glc-6-P and Fru-6-P paralleled each other and exhibited a hyperbolic pattern towards a new elevated steady-state signal level, perhaps characteristic of low temperature metabolic necessity. Hexose phosphates are immediate precursors for sucrose biosynthesis and, indeed, it can be seen that their increase was kinetically well matched to the $T_{0.5[MAX]}$ of sucrose (Figure 7c).

Sucrose signal intensity increased over 1–4 h of CA with an estimated $T_{0.5[MAX]}$ of about 1 h. This increase was followed by signal increases in its breakdown products Glc and Fru (Figures. 5b, 7c) with $T_{0.5[MAX]}$ values of about 1 h. Given the nearly 1:1 hyperbolic pattern, the observed increases in Glc and Fru suggest a common source and most likely reflect the breakdown of sucrose by invertase. Transcript signal levels for genes encoding invertases remained constant and were not affected by CA (data not shown). Therefore, the increased level of substrate (sucrose) and compartmentation during CA would be the driving factor that resulted in the parallel and hyperbolic increases in products Glc and Fru. In contrast, galactinol, following the observed increase in transcript signal levels for galactinol synthase (Figure 6b), accumulated at relatively later stages

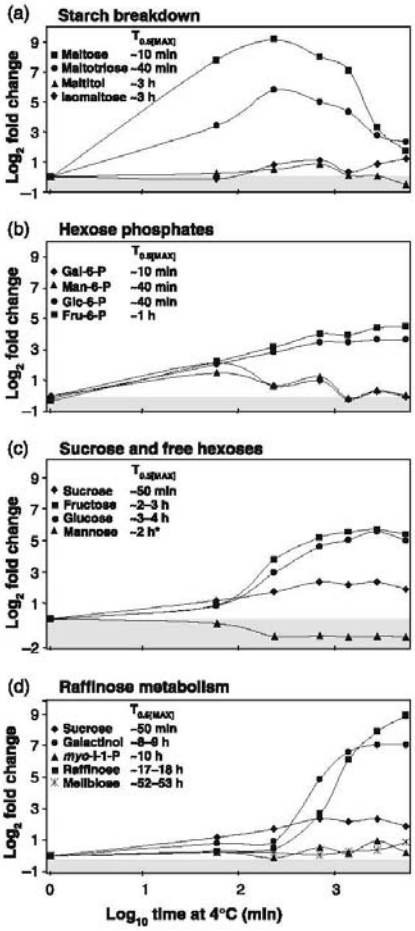


Figure 7. Time course of carbohydrate accumulation during cold acclimation. (a) Starch degradation products, (b) hexose phosphates, (c) sucrose and free hexoses, and (d) raffinose metabolites. $T_{0.5[MAX]}$ time-point estimates are derived from interpolation of the individual metabolite time-course profiles where time points are calculated in which each metabolite reached 50% of its maximum (*minimum) concentration during the time course. The shaded area indicates pre-stress steady-state levels and below. For metabolites whose levels declined, the $T_{0.5[MIN]}$ are indicated with an asterisk.

($T_{0.5[MAX]}$ = 8–9 h) during CA (Figure 7d). Finally, the combined accretion of galactinol together with the earlier rise in sucrose contribute to the robust biosynthesis and build-up of raffinose at sequentially later stages of CA ($T_{0.5[MAX]}$ = 17–18 h; Figure 7d). Thus, overall, the estimated $T_{0.5[MAX]}$

 The Plant Journal, (2007), **50**, 967–981

values suggest an integrated metabolic sequence of events resulting in the apparent accumulation of carbohydrates during CA that is summarized in Figure 7.

Discussion

It is well established that CA leading to enhanced freezing tolerance is a complex process involving coordinated activation of many biochemical pathways (Guy, 1990, 1999; Shinozaki *et al.*, 2003; Stitt and Hurry, 2002; Sung *et al.*, 2003; Thomashow, 1999). However, only recent metabolite profiling studies have revealed the degree to which the metabolome of Arabidopsis is altered in response to low temperatures, the multiplicity of the compatible solute-like network and the metabolic pathways involved (Browse and Lange, 2004; Cook *et al.*, 2004; Gray and Heath, 2005; Hannah *et al.*, 2006; Kaplan *et al.*, 2004). In this study, transcript and metabolite profiling analyses were combined in order to achieve a more comprehensive understanding of the dynamic sequence of events linking gene-to-metabolite networks during CA and adaptation to low temperatures. The present findings complement and extend previous research by others, where ectopic overexpression of transcriptional activators linked to cold-regulated gene expression have demonstrated a 'CA-like' state with regards to molecular, biochemical and metabolic alterations in plants at non-inductive temperatures (Cook *et al.*, 2004; Gilmour *et al.*, 2000). In the present study, PCA clearly shows that the global pattern of changes in transcript and metabolite signal levels during the time course of CA exhibit a comparable progression, and suggests an integrated linkage between molecular and metabolic processes on a broad scale (Figure 1). We also find that with respect to gene expression patterns, changes in metabolite signal levels for amino acids, particularly Pro, and some soluble sugars do not necessarily tightly conform to transcript abundance responses. Such correlative disconnects signify that both transcript abundance dependent and independent regulatory processes are important factors in metabolic and cellular acclimation to low temperatures (Table 1). This may be the reason why global changes in metabolites may not correlate well with the ability to acclimate (Hannah *et al.*, 2006). By combining transcriptome and metabolome responses, it is possible to reveal in a case-by-case manner whether a given metabolic process is primarily regulated by transcript abundance and/or by other regulatory mechanisms downstream of steady-state mRNA levels (Smith *et al.*, 2004; Ter Kuile and Westerhoff, 2001), and when taken together can provide new insight to help define additional regulatory control points of biochemical pathways. Given the overall tight congruence of the transcriptome and metabolome variance progression over the time course of CA, it is logical to consider that these two organizational levels of plant cells are not unidirectionally linked in a 'central dogma' sense, but are engaged in a feedforward/feedback circuit (Moore *et al.*, 2003; Rolland *et al.*, 2002) that continuously integrates gene expression outputs with metabolic outputs and status, and vice versa in concert with prevailing environmental conditions. Both GABA and Pro metabolism (Figure 4) illustrates primary and secondary responses during CA that are consistent with bidirectional regulatory control.

Proline, GABA and free sugars are stress-related metabolites that share compatible solute-like properties, accumulate at low temperatures, and have been shown to have a role in stabilizing and protecting proteins and membranes from freezing damage (Carpenter and Crowe, 1988; Heber *et al.*, 1971; Strauss and Hauser, 1986; Uemura *et al.*, 2003; Wanner and Junttila, 1999; Yancey *et al.*, 1982). Proline accumulation is a common occurrence in plants in response to abiotic stress. In stressed as well as non-stressed plants, Pro accumulation follows the coordinate activation of biosynthesis and inactivation of its degradation; and in non-stressed plants, decreases in Pro levels are caused by coordinate inhibition of biosynthesis and by activation of degradation (Kiyosue *et al.*, 1996; Peng *et al.*, 1996).

Proline levels during water deficit and salinity stress have been linked with changes in gene expression for some of the enzymes of proline metabolism in an integrated transcriptome and metabolome study with grape (Cramer *et al.*, 2007). In the present study, an early increase in Pro signal levels after 4 h of CA was observed, whereas *P5CS* transcript signal levels increased only much later, after 12 h of CA (Figure 3b). It would seem reasonable that a large reduction of protein synthesis without an equally large reduction in amino acid biosynthesis could account for this immediate rise in Pro content. However, during the first 4 h of CA, Glu levels were largely unchanged, so the increased Pro cannot result from high Glu substrate levels. Thus, the rapid rise in the Pro signal levels in the absence of changes in transcript signal levels suggests that low temperature somehow influences the activity or activation state of *P5CS*, perhaps by a desensitization of feedback regulation. During intermediate and later stages of CA (12–96 h), *P5CS* transcript signal levels decreased whereas *PDH* transcript signal levels increased. Throughout this time, Pro signal intensity continued to increase (Figure 3b). This suggests that Pro synthesis continued to exceed utilization and degradation well after the peak of gene expression for *P5CS*, and well after increased *PDH* expression. Thus, Pro metabolism provides an example of transcript abundance independent regulation of an important biochemical pathway during the time course of CA. The synthesis of Pro from Glu and its degradation leading to the resynthesis of Glu from P5C during CA constitutes a potentially futile cycle, which would consume ATP and reducing potential if it were not for the fact of spatial separation where Pro synthesis is cytosolic and degradation is mitochondrial. Whenever a potential

futile cycle could exist, post-transcriptional regulatory mechanisms act to prevent such a cycle from occurring. One way Pro levels might continue to rise as *PDH* expression is induced is to regulate the transport of Pro from the cytosol into the mitochondria. Other forms of regulation of proline metabolism are known. For example, *P5CS* and P5CR activities are subject to feedback inhibition by Pro under stress and non-stress conditions, whereas *P5CDH* expression is induced by Pro (Deuschle *et al.*, 2001; Hong *et al.*, 2000). In the present study, at the steady-state transcript signal level, *P5CDH* was only very slightly induced during CA when Pro signal levels were robustly rising. Recently, a novel mode of post-transcriptional regulation of Pro metabolism was discovered where an antisense overlapping gene pair of *P5CDH*, and a stress-related gene of unknown function, *SRO5*, gives rise to two types of siRNAs that initiate *P5CDH* mRNA cleavage (Borsani *et al.*, 2005). Regarding *P5CR*, increased transcript levels during stress have been shown to result from increased mRNA stability, whereas translation initiation was also inhibited and protein levels remained unchanged (Hua *et al.*, 2001). Therefore, during CA, in addition to compartmentation, it appears probable that several forms of post-transcriptional and/or post-translational regulation are operational in Pro metabolism.

GABA metabolism, in contrast with that of Pro, provides an example of a transcript abundance dependent response during CA. An increase in *GAD* transcript signal levels was detected after 12 h of CA, which was followed by a later increase in GABA signal after 24 h (Figure 4b). Subsequently, the decline in GABA signal levels after 24 h was accompanied by a decrease in *GAD* transcript signal levels and a heightened signal level in *SSADH* transcript. Even in the presence of high Glu signal levels, the steady-state levels of GABA appear to result from a balance of synthesis and catabolism. Taken together, these coordinate changes reveal that GABA catabolism is, in part, a function of control of transcript abundance, possibly linked to a metabolite feedback mechanism.

A novel finding that emerged from the current study was the apparent sequential changes in steady-state patterns for metabolites involved in sugar accumulation during CA (Figure 7). Metabolic adjustments often occur on very short timescales. Yet, in the present time-course study taking place over intervals of hours and days, it seems rather remarkable to be able to record the metabolic progression beginning with starch breakdown products and ending with raffinose breakdown products.

The integration of transcript and metabolite profiling data allowed the detection of potential rate-limiting steps and key regulatory enzymes of sugar biosynthesis pathways during CA. For example, transcript abundance regulation of *HEX* and *FK* during the early stages of CA probably contributes to the early increases in hexose phosphate pools of Glc-6-P and Fru-6-P, which are precursors of sucrose biosynthesis, leading to enhanced sucrose biosynthesis upon CA. On the other hand, *SPS* and *SPP* transcript signal levels either increased much later or were not significantly affected by CA, and, therefore, transcript abundance dependent processes for these enzymes do not seem to play major regulatory roles in the initial rise in sucrose at low temperatures (Figure 5). SPS light/dark activity is regulated by phosphorylation/dephosphorylation, in part by an SNF1-related protein kinase [PK(III)] (Toroser *et al.*, 2000) that is inhibited by Glc-6-P. The rapid rise in Glc-6-P content would result in a reduction of phosphorylation and rapid activation of SPS activity at low temperature without an increase in *SPS* gene expression. During the later stages of CA, the increases in transcript signal levels of both galactinol synthase and raffinose synthase preceded later increases in galactinol and raffinose in a fashion that reflects the fact that transcript abundance acts to regulate these genes, which are essential for the induction of galactinol and raffinose biosynthesis (Figure 6).

Recently, it has been demonstrated that a key element in maintaining flux through carbohydrate metabolism in the cold is to control the partitioning of metabolites between the chloroplast and the cytosol. Arabidopsis appears to modulate the expression of different metabolite transporters to maintain a balanced carbon flow during CA and low-temperature plant development (Lundmark *et al.*, 2006). Therefore, the functioning of metabolite transporters is expected to have a major influence on carbohydrate metabolic progression during CA.

For abundant metabolites, non-targeted metabolite profiling has the potential to capture information about relative levels of substrates and products for reversible rate-controlling enzymatic reactions, which can provide new insight into *in vivo* metabolic conditions. To determine whether it is possible to infer relative quantitative relationships between substrates and products based on metabolite profiling signal intensity data, we examined the reversible reaction catalyzed by raffinose synthase. Substrates of this enzyme are galactinol and sucrose and the products are raffinose and *myo*-inositol, and our dataset contained all of these metabolites. Raffinose synthase has been purified from pea (Peterbauer *et al.*, 2002) and its kinetic properties characterized. The K_{eq} for the pea enzyme was determined to be 4.1. We used the steady-state signal values from metabolite profiling to determine an instantaneous estimate of the product/substrate ratio, which when compared with the K_{eq} = 4.1 would reflect the magnitude of divergence from equilibrium. We plotted the $\log_2$ ratio of the instantaneous/K_{eq} values (Figure 8). If the resultant values are greater than zero, the reaction is predicted to proceed towards substrates; if the values are less than zero, the reaction is predicted to proceed towards products. Prior to transfer to low temperature the value was positive, and the expected direction of the reaction would be towards substrates. By 4 h

 The Plant Journal, (2007), **50**, 967–981

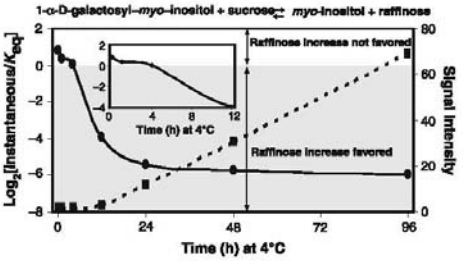


Figure 8. The association of metabolite relative signal values to raffinose synthase equilibrium constant K_{eq} and raffinose accumulation during cold acclimation.
The K_{eq} for pea raffinose synthase is 4.1 (Peterbauer *et al.*, 2002). The inset shows an expanded time course change from 0 to 12 h. The shaded area indicates where the Log_2[Instantaneous/K_{eq}] ratio (•) favors raffinose synthesis. Raffinose signal intensity is indicated by (■).

of CA the ratio was zero. At all time-points after 4 h, the ratio values were strongly negative indicating that the reaction would proceed in the direction of products. When raffinose signal levels are plotted, it can be easily seen that its rise is matched perfectly with the prediction of relative steady-state levels of products to substrates in the presence of the enzyme. This result is in exact agreement with the previous finding that raffinose accumulation is controlled by the presence of galactinol synthase activity, and by the levels of the substrates galactinol and sucrose (Karner *et al.*, 2004). Thus, we demonstrate that it is possible for metabolite profiling signal output data to complement known biochemical characteristics of enzymes, and to provide a more meaningful picture of the steady-state status of metabolic processes.

In conclusion, integrative analysis of transcript and metabolite profiling provides a powerful approach to better understand gene-to-metabolite networks in biological systems, and for the further characterization of key regulatory enzymes and metabolites involved in plant acclimation to environmental signals (Buckhout and Thimm, 2003; Friedman and Pichersky, 2005; Oksman-Caldentey and Saito, 2005). Taken together, the findings obtained from this combined transcript and metabolite profiling analysis illustrate the integration and dynamism, with regard to changes within selected major biochemical pathways activated during CA in Arabidopsis. The present findings also reveal metabolite kinetic patterns that reflect transcript abundance independent regulatory mechanisms and crucial control points of regulation. In addition, our findings reveal that for many metabolic adjustments that occur during CA, additional regulatory mechanisms downstream of mRNA steady-state levels are operational. Ultimately, the complete dissection of the complex process of CA through systems biology approaches will lead to the deconvolution of the progressive acquisition of stress tolerance into component molecular, biochemical, metabolic and physiologic contributors.

Experimental procedures

Plant growth and cold acclimation

Arabidopsis (*Arabidopsis thaliana*, ecotype Columbia) plants were grown using the 'deli pot' culture system (Sung and Guy, 2003) in controlled environment chambers for 3 weeks at 20°C with a photoperiod cycle of 15-h light/9-h dark. Irradiance was provided with incandescent bulbs and cool white fluorescent tubes, and ranged between 100 and 130 μmol $m^{-2}s^{-1}$ at canopy height. Plants were at the eight-leaf stage of development when the experiment was begun. The details of the experimental conditions have been previously described (Kaplan *et al.*, 2004). Briefly, the CA treatment was initiated 2 h after the onset of the light period, which allowed the harvest of all samples within the light period, by placing the plants at 4°C. Samples were taken at 1, 4, 12, 24, 48 and 96 h of CA. Additionally, untreated controls were taken at time zero (2 h into the light period) and at 4 h (6 h into the light period) after the experiment began. All samples were rapidly harvested, flash-frozen in liquid nitrogen (<30 sec), and stored at −80°C until RNA and metabolite extraction.

Transcript profiling

Transcript profiling followed the Miame protocol recommendations (http://www.affymetrix.com/products/arrays/specific/arab.affx). Total RNA for each sample was extracted from the aerial tissues of Arabidopsis plants using the QIAGEN RNeasy® Plant Mini Kit (Qiagen, http://www1.qiagen.com) according to the manufacturer's instructions. The RNA samples were further prepared for hybridizations according to the protocols outlined in the GeneChip® Expression Analysis Technical Manual (Affymetrix, http://www.affymetrix.com). The samples were hybridized to the Affymetrix Arabidopsis ATH1 GeneChip (Affymetrix). The experiment was conducted three times yielding three replicate measurements from separate RNA extractions per time point. Hybridizations were performed at the UFSCC/ICBR Microarray Core facility at the University of Florida, Gainesville, FL, USA. One-way ANOVA tests were performed on signal values to identify probe sets exhibiting significant changes in signal levels at $P \leq 0.05$, and the Kruskal–Wallis test was carried out for the pairwise comparisons.

The results from the microarray experiments have been deposited to the NASCArrays (http://affymetrix.arabidopsis.info/donating.html) according to MIAME guidelines under the accession number NASCARRAYS-404.

Metabolite time-course plots in Figures 3, 4, 5 and 6 were made using data extracted from Table S3 of Kaplan *et al.* (2004).

Data for selected metabolite profiles that were presented in Figure 4 of Kaplan *et al.* (2004) are replotted as Log_2 ratio of control to treatment signal values. The ratio of control to treatment metabolite signal values for proline, GABA, fructose, glucose, F-6-P, G-6-P, sucrose, galactinol and raffinose can be found in Table S3 of Kaplan *et al.* (2004).

Principal component analysis

Parallel PCA for transcripts and metabolites was performed with the S-PLUS 2000 software package, standard edition release 3 (Insightful, http://www.insightful.com) on log_{10}-transformed relative

responses, $\log_{10}(R_i)$. The denominator of the quotient R_i was the average response of non-treated control samples at time zero ($R_i = N_i \times \text{avg}N_{t0}^{-1}$). Responses ($N_i$) were volume corrected and normalized by the fresh weight of each sample. Log-transformed relative responses were averaged at each time point. Metabolites or genes with more than one missing time point or low overall variance were excluded. PCA of the top 1117 or 5400 most variable genes exhibited highly similar results.

Databases for metabolite pathways

AraCyc metabolic pathways databases were downloaded from The Arabidopsis Information Resource (TAIR) (http://ftp://ftp.arabidopsis.org/home/tair/Pathways). The predicted subcellular localization of the various enzymes was indicated according to the TAIR website (http://ftp://ftp.arabidopsis.org/home/tair/Proteins/Properties).

Acknowledgements

We thank Cameron Schiller, Nicole Gatzke, Dale Haskell and Kil-Jae Lee for their help with this work. We also thank Dr Lothar Willmitzer and the Max Planck Society for continuing support. This research was supported by grants from NASA #NAG10-316, USDA NRI #2000-35100-9532 and #2002-35100-12110 and the Institute of Food & Agricultural Sciences at the University of Florida.

Supplementary material

The following supplementary material is available for this article online:

Table S1 Probe sets showing statistically different signal intensity during cold acclimation.

Table S2 Gene expression for cold acclimation responsive pathways.

Table S3 Metabolites for cold acclimation responsive pathways.

Appendix S1 Methods for microarray and metabolite profiling.

This material is available as part of the online article from http://www.blackwell-synergy.com.

References

Bohnert, H.J., Gong, Q., Li, P. and Ma, S. (2006) Unraveling abiotic stress tolerance mechanisms – getting genomics going. *Curr. Opin. Plant Biol.* **9**, 180–188.

Borsani, O., Zhu, J., Verslues, P.E., Sunkar, R. and Zhu, J.K. (2005) Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. *Cell*, **123**, 1279–1291.

Bouche, N. and Fromm, H. (2004) GABA in plants: just a metabolite? *Trends Plant Sci.* **9**, 110–115.

Bouche, N., Fait, A., Bouchez, D., Moller, S.G. and Fromm, H. (2003) Mitochondrial succinic-semialdehyde dehydrogenase of the gamma-aminobutyrate shunt is required to restrict levels of reactive oxygen intermediates in plants. *Proc. Natl Acad. Sci. USA*, **100**, 6843–6848.

Browse, J. and Lange, B.M. (2004) Counting the cost of a cold-blooded life: metabolomics of cold acclimation. *Proc. Natl Acad. Sci. USA*, **101**, 14996–14997.

Buckhout, T.J. and Thimm, O. (2003) Insights into metabolism obtained from microarray analysis. *Curr. Opin. Plant Biol.* **6**, 288–296.

Carpenter, J.F. and Crowe, J.H. (1988) The mechanism of cryoprotection of proteins by solutes. *Cryobiology*, **25**, 244–255.

Cook, D., Fowler, S., Fiehn, O. and Thomashow, M.F. (2004) A prominent role of the CBF cold response pathway in configuring the low temperature metabolome of Arabidopsis. *Proc. Natl Acad. Sci. USA*, **101**, 15243–15248.

Coruzzi, G. and Last, R. (2000) Amino acids. In *Biochemistry and Molecular Biology of Plants*. (Buchanan, B.B., Wilhelm, G. and Jones, R.L., eds). Waldorf, MD: American Society of Plant Biologists, pp. 358–410.

Cramer, G.R., Ergul, A., Grimplet, J. *et al.* (2007) Water and salinity stress in grapevines: early and late changes in transcript and metabolite profiles. *Funct. Intger. Genomics*, **7**, 111–134.

Dennis, D.T. and Blakeley, S.D. (2000) Carbohydrate metabolism. In *Biochemistry and Molecular Biology of Plants*. (Buchanan, B.B., Wilhelm, G. and Jones, R.L., eds.). Waldorf, MD: American Society of Plant Biologists, pp. 630–675.

Deuschle, K., Funck, D., Hellmann, H., Daschner, K., Binder, S. and Frommer, W.B. (2001) A nuclear gene encoding mitochondrial Delta-pyrroline-5-carboxylate dehydrogenase and its potential role in protection from proline toxicity. *Plant J.* **27**, 345–356.

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R.N. and Wilmitzer, L. (2000) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* **18**, 1157–1161.

Fowler, S. and Thomashow, M.F. (2002) Arabidopsis transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway. *Plant Cell*, **14**, 1675–1690.

Friedman, E. and Pichersky, E. (2005) Metabolomics, genomics, proteomics, and the identification of enzymes and their substrates and products. *Curr. Opin. Plant Biol.* **8**, 242–248.

Gilmour, S.J., Sebolt, A.M., Salazar, M.P., Everard, J.D. and Thomashow, M.F. (2000) Overexpression of the Arabidopsis CBF3 transcriptional activator mimics multiple biochemical changes associated with cold acclimation. *Plant Physiol.* **124**, 1854–1865.

Gray, G.R. and Heath, D. (2005) A global reorganization of the metabolome in Arabidopsis during cold acclimation is revealed by metabolomic fingerprinting. *Physiol. Plant.* **124**, 236–248.

Guy, C.L. (1990) Cold acclimation and freezing stress tolerance: role of protein metabolism. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **41**, 187–223.

Guy, C.L. (1999) Molecular responses of plants to cold shock and cold acclimation. *J. Mol. Microbiol. Biotechnol.* **12**, 231–242.

Guy, C.L., Huber, J.L.A. and Huber, S.C. (1992) Sucrose phosphate synthase and sucrose accumulation at low temperature. *Plant Physiol.* **100**, 502–508.

Hannah, M.A., Wiese, D., Freund, S., Fiehn, O., Heyer, A.G. and Hincha, D.K. (2006) Natural genetic variation of freezing tolerance in Arabidopsis. *Plant Physiol.* **142**, 98–112.

Hayashi, F., Ichino, T., Osanai, M. and Wada, K. (1996) Oscillation and regulation of proline content by P5CS and ProDH gene expressions in the light/dark cycles in *Arabidopsis thaliana* L. *Plant Cell Physiol.* **41**, 1096–1101.

Heber, E., Tyankova, L. and Santarius, K.A. (1971) Stabilization and inactivation of biological membranes during freezing in the presence of amino acids. *Biochim. Biophys. Acta.* **241**, 578–592.

Hirai, M.Y., Yano, M., Goodenowe, D.B., Kanaya, S., Kimura, T., Awazuhara, M., Arita, M., Fujiwara, T. and Saito, K. (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA*, **101**, 10205–10210.

Hirai, M.Y., Klein, M., Fujikawa, Y. *et al.* (2005) Elucidation of gene-to-gene and metabolite-to-gene networks in Arabidopsis by

integration of metabolomics and transcriptomics. *J. Biol. Chem.* **280**, 25590–25595.

Hong, Z., Lakkineni, K., Zhang, Z. and Verma, D.P. (2000) Removal of feedback inhibition of delta(1)-pyrroline-5-carboxylate synthetase results in increased proline accumulation and protection of plants from osmotic stress. *Plant Physiol.* **122**, 1129–1136.

Hu, C.A., Delauney, A.J. and Verma, D.P. (1992) A bifunctional enzyme (delta 1-pyrroline-5-carboxylate synthetase) catalyzes the first two steps in proline biosynthesis in plants. *Proc. Natl Acad. Sci. USA*, **89**, 9354–9358.

Hua, X.J., Van de Cotte, B., Van Montagu, M. and Verbruggen, N. (2001) The 5′ untranslated region of the At-P5R gene is involved in both transcriptional and post-transcriptional regulation. *Plant J.* **26**, 157–1169.

Kaplan, F., Kopka, J., Dale, W.H., Zhao, W., Schiller, K.C., Gatzke, N., Sung, D.Y. and Guy, C.L. (2004) Exploring the temperature-stress metabolome of Arabidopsis. *Plant Physiol.* **136**, 4159–4168.

Karner, U., Peterbauer, T., Raboy, V., Jones, D.A., Hedley, C.L. and Richter, A. (2004) *myo*-Inositol and sucrose concentrations affect the accumulation of raffinose family oligosaccharides in seeds. *J. Exp. Bot.* **55**, 1981–1987.

Kirschner, M.W. (2005) The meaning of systems biology. *Cell*, **121**, 503–504.

Kiyosue, T., Yoshiba, Y., Yamaguchi-Shinozaki, K. and Shinozaki, K. (1996) A nuclear gene encoding mitochondrial proline dehydrogenase, an enzyme involved in proline metabolism, is upregulated by proline but downregulated by dehydration in Arabidopsis. *Plant Cell*, **8**, 1323–1335.

Kreps, J.A., Wu, Y., Chang, H.-S., Zhu, T., Wang, X. and Harper, J.F. (2002) Transcriptome changes for Arabidopsis in response to salt, osmotic and cold stresses. *Plant Physiol.* **130**, 2129–2141.

Levitt, J. (1972) *Responses on Plants to Environmental Stresses.* New York, NY: Academic Press.

Lundmark, M., Cavaco, A.M., Trevanion, S. and Hurry, V. (2006) Carbon partitioning and export in transgenic Arabidopsis thaliana with altered capacity for sucrose synthesis grown at low temperature: a role for metabolite transporters. *Plant Cell Environ.* **29**, 1703–1714.

Moore, B., Zhou, L., Rolland, F., Hall, Q., Cheng, W.-H., Liu, Y.-X., Hwang, I., Jones, T. and Sheen, J. (2003) Role of the Arabidopsis sucrose sensor HXK1 in nutrient, light, and hormonal signaling. *Science*, **300**, 332–336.

Oksman-Caldentey, K.M. and Saito, K. (2005) Integrating genomics and metabolomics for engineering plant metabolic pathways. *Curr. Opin. Biotechnol.* **16**, 174–179.

Oksman-Caldentey, K.M., Inze, D. and Oresic, M. (2004) Connecting genes to metabolites by a systems biology approach. *Proc. Natl Acad. Sci. USA*, **101**, 9949–9950.

Peng, Z., Lu, Q. and Verma, D.P. (1996) Reciprocal regulation of delta 1-pyrroline-5-carboxylate synthetase and proline dehydrogenase genes controls proline levels during and after osmotic stress in plants. *Mol. Gen. Genet.* **253**, 334–341.

Peterbauer, T., Mach, L., Mucha, J. and Richter, A. (2002) Functional expression of a cDNA encoding pea (*Pisum sativum* L.) raffinose synthase, partial purification of the enzyme from maturing seeds, and steady-state kinetic analysis of raffinose synthesis. *Planta*, **215**, 839–846.

Renaut, J., Hausman, J.F. and Wisniewski, M.E. (2006) Proteomics and low temperature studies: bridging the gap between gene expression and metabolism. *Physiol. Plant.* **126**, 97–109.

Rohde, P., Hincha, D.K. and Heyer, A.G. (2004) Heterosis in the freezing tolerance of crosses between two *Arabidopsis thaliana* accessions (Colombia-0 and C24) that show differences in non-acclimated and acclimated freezing tolerance. *Plant J.* **38**, 790–799.

Rolland, F., Moore, B. and Sheen, J. (2002) Sugar sensing and signaling in plants. *Plant Cell*, **14**, 185–205.

Scholz, M., Kaplan, F., Guy, C.L., Kopka, J. and Selbig, J. (2005) Non-linear PCA: a missing data approach. *Bioinformatics*, **21**, 3887–3895.

Seki, M., Narusaka, M., Ishida, J. *et al.* (2002) Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. *Plant Cell*, **31**, 279–291.

Shelp, B.J., Bown, A.W. and McLean, M.D. (1999) Metabolism and functions of gamma-aminobutyric acid. *Trends Plant Sci.* **4**, 446–452.

Shinozaki, K. and Dennis, E.S. (2003) Cell signaling and gene regulation: global analyses of signal transduction and gene expression profiles. *Curr. Opin. Plant Biol.* **6**, 405–409.

Shinozaki, K., Yamaguchi-Shinozaki, K. and Seki, M. (2003) Regulatory network of gene expression in the drought and cold stress responses. *Curr. Opin. Plant Biol.* **6**, 410–417.

Smith, S.M., Fulton, D.C., Chia, T., Thorneycroft, D., Chapple, A., Dunstan, H., Hylton, C., Zeeman, S.C. and Smith, A.M. (2004) Diurnal changes in the transcriptome encoding enzymes of starch metabolism provide evidence for both transcriptional and post-transcriptional regulation of starch metabolism in Arabidopsis leaves. *Plant Physiol.* **136**, 2687–2699.

Stitt, M. and Hurry, V. (2002) A plant for all seasons: alterations in photosynthetic carbon metabolism during cold acclimation in Arabidopsis. *Curr. Opin. Plant Biol.* **5**, 199–206.

Strauss, G. and Hauser, H. (1986) Stabilization of lipid bilayer vesicles by sucrose during freezing. *Proc. Natl Acad. Sci. USA*, **83**, 2422–2426.

Strizhov, N., Abraham, E., Okresz, L., Blicking, S., Zilberstein, A., Schell, J., Koncz, C. and Szabados, L. (1997) Differential expression of two *P5CS* genes controlling proline accumulation during salt-stress requires ABA and is regulated by *ABA1*. ABI1 and AXR2 in Arabidopsis. *Plant J.* **12**, 557–569.

Sung, D.Y. and Guy, C.L. (2003) Physiological and molecular assessment of altered expression of Hsc70-1 in Arabidopsis. Evidence for pleiotropic consequences. *Plant Physiol.* **132**, 979–987.

Sung, D.Y., Kaplan, F., Lee, K.J. and Guy, C.L. (2003) Acquired tolerance to temperature extremes. *Trends Plant Sci.* **8**, 179–187.

Taji, T., Ohsumi, C., Iuchi, S., Seki, M., Kasuga, M., Kobayashi, M., Yamaguchi-Shinozaki, K. and Shinozaki, K. (2002) Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance of *Arabidopsis thaliana*. *Plant J.* **29**, 417–426.

Ter Kuile, B.H. and Westerhoff, H.V. (2001) Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett.* **500**, 169–171.

Thomashow, M. (1999) Plant cold acclimation: freezing tolerance genes and regulatory mechanisms. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **50**, 571–599.

Tohge, T., Nishiyama, Y., Hirai, M.Y. *et al.* (2005) Functional genomics by integrated analysis of metabolome and transcriptome of Arabidopsis plants over-expressing an MYB transcription factor. *Plant J.* **42**, 218–235.

Toroser, D., Plaut, Z. and Huber, S.C. (2000) Regulation of a plant SNF1-related protein kinase by glucose-6-phosphate. *Plant Physiol.* **123**, 403–412.

Uemura, M., Warren, G. and Steponkus, P.L. (2003) Freezing sensitivity in the SFR4 mutant of Arabidopsis is due to low sugar content and is manifested by loss of osmotic responsiveness. *Plant Physiol.* **131**, 1800–1807.

Verbruggen, N., Villarroel, R. and Van Montagu, M. (1993) Osmoregulation of a pyrroline-5-carboxylate reductase gene in *Arabidopsis thaliana*. *Plant Physiol.* **103**, 771–781.

Vogel, J.T., Zarka, D.G., Van Buskirk, H.A., Fowler, S.G. and Thomashow, M.F. (2005) Roles of the CBF2 and ZAT12 transcription factors in configuring the low temperature transcriptome of Arabidopsis. *Plant J.* **41**, 195–211.

Wang, X., Li, W. and Welti, R. (2006) Profiling lipid changes in plant response to low temperatures. *Physiol. Plant.* **126**, 90–96.

Wanner, L.A. and Junttila, O. (1999) Cold-induced freezing tolerance in Arabidopsis. *Plant Physiol.* **120**, 391–399.

Xin, Z. and Browse, J. (1998) *Eskimo1* mutants of Arabidopsis are constitutively freezing-tolerant. *Proc. Natl Acad. Sci. USA*, **95**, 7799–7804.

Yancey, P.H., Clark, M.E., Hand, S.C., Bowlus, R.D. and Somero, G.N. (1982) Living with water stress: evolution of osmolyte systems. *Science*, **217**, 1214–1222.

Yoshiba, Y., Kiyosue, T., Katagiri, T., Ueda, H., Mizoguchi, T., Yamaguchi-Shinozaki, K., Wada, K., Harada, Y. and Shinozaki, K. (1995) Correlation between the induction of a gene for delta 1-pyrroline-5-carboxylate synthetase and the accumulation of proline in *Arabidopsis thaliana* under osmotic stress. *Plant J.* **7**, 751–760.

 The Plant Journal, (2007), **50**, 967–981

The Plant Journal (2004), **39**, 487–512 doi: 10.1111/j.1365-313X.2004.02150.x

Global changes in transcription orchestrate metabolic differentiation during symbiotic nitrogen fixation in *Lotus japonicus*

Gillian Colebatch[†,§], Guilhem Desbrosses[§], Thomas Ott, Lene Krusell, Ombretta Montanari, Sebastian Kloska[‡], Joachim Kopka and Michael K. Udvardi[*]

Max Planck Institute for Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm, Germany

Received 30 January 2004; revised 7 May 2004; accepted 10 May 2004.
[*]For correspondence (fax +49 331 567 8250; e-mail udvardi@mpimp-golm.mpg.de).
[†]Present address: CSIRO Division of Entomology, GPO Box 1700, Canberra, ACT 2601, Australia.
[‡]Present address: Scienion AG, Volmerstr. 7a, 12489 Berlin, Germany.
[§]These authors contributed equally to this work.

Summary

Research on legume nodule metabolism has contributed greatly to our knowledge of primary carbon and nitrogen metabolism in plants in general, and in symbiotic nitrogen fixation in particular. However, most previous studies focused on one or a few genes/enzymes involved in selected metabolic pathways in many different legume species. We utilized the tools of transcriptomics and metabolomics to obtain an unprecedented overview of the metabolic differentiation that results from nodule development in the model legume, *Lotus japonicus*. Using an array of more than 5000 nodule cDNA clones, representing 2500 different genes, we identified approximately 860 genes that were more highly expressed in nodules than in roots. One-third of these are involved in metabolism and transport, and over 100 encode proteins that are likely to be involved in signalling, or regulation of gene expression at the transcriptional or post-transcriptional level. Several metabolic pathways appeared to be co-ordinately upregulated in nodules, including glycolysis, CO_2 fixation, amino acid biosynthesis, and purine, haem, and redox metabolism. Insight into the physiological conditions that prevail within nodules was obtained from specific sets of induced genes. In addition to the expected signs of hypoxia, numerous indications were obtained that nodule cells also experience P-limitation and osmotic stress. Several potential regulators of these stress responses were identified. Metabolite profiling by gas chromatography coupled to mass spectrometry revealed a distinct metabolic phenotype for nodules that reflected the global changes in metabolism inferred from transcriptome analysis.

Keywords: legume, nodule, transcriptome, metabolome, symbiotic nitrogen fixation, *Lotus japonicus*.

Introduction

Biological nitrogen fixation, which reduces N_2 to ammonium, is the largest source of available nitrogen for life on earth (Newton, 2000). Much of this ammonium comes from symbiotic nitrogen fixation (SNF) by rhizobia within legume root nodules. The Leguminoseae is one of the most successful families of land plants, in part because of SNF, which enables legumes to colonize soils that contain little or no available nitrogen. This feature, together with the nutritious, protein-rich seeds that they produce, placed legumes at the origins of ancient agriculture. To this day, legumes remain an essential part of traditional and modern agriculture.

Legume–rhizobia symbioses are beneficial to both partners. In exchange for a generous supply of reduced nitrogen from rhizobia, the plant provides its micro-symbionts with reduced carbon and all other necessary nutrients (Udvardi and Day, 1997). Such mutualism requires exquisite integration of plant and bacterial metabolism to avoid exploitation of one partner by the other (Lodwig *et al.*, 2003), although how this is achieved at a biochemical level remains unclear.

Establishment of an effective, nitrogen-fixing symbiosis between legumes and rhizobia is a complex process, involving signalling and recognition by both partners from the outset (Long, 2001; Stougaard, 2001). Rhizobial attachment to root hairs, penetration of the epidermis, and invasion of cortical tissue via the infection thread are accompanied by initiation of meristematic activity in root

cortical and pericycle cells and suppression of plant defence responses. Release of rhizobia from infection threads into individual cortical cells is achieved via endocytosis, which leaves the rhizobia enclosed in a plant membrane called the peribacteroid or symbiosome membrane (SM) that isolates them from the host cell cytoplasm. The resulting organelle is called a symbiosome. Rhizobia continue to divide until infected cells are packed with thousands of bacteria, which are surrounded, either individually or in small groups, by the SM. Other profound changes occur during nodule development, including the establishment of a vascular network that delivers photosynthate, mostly in the form of sucrose, to the nodule tissues, and which also facilitates the export of nitrogen-containing compounds from active nodules. Before nitrogen fixation can take place, however, a micro-aerobic environment is established inside nodules, which triggers differentiation of rhizobia into nitrogen-fixing bacteroids (Batut and Boistard, 1994; Fischer, 1996). Over the past 20 years, significant insights into many aspects of SNF have been gained. Critical early signalling events have been uncovered, numerous bacterial genes essential for nodule development or function have been identified, and important aspects of bacteroid and legume nodule metabolism have been elucidated (Day and Copeland, 1991; Denarie *et al.*, 1996; Downie and Walker, 1999; Kahn *et al.*, 1998; Long, 2001; Stougaard, 2000; Udvardi and Day, 1997; Vance *et al.*, 1994). However, few plant genes essential for normal nodule development (Endre *et al.*, 2002; Krusell *et al.*, 2002; Nishimura *et al.*, 2002; Schauser *et al.*, 1999; Searle *et al.*, 2003; Stracke *et al.*, 2002) and only one that is crucial for symbiotic metabolism in mature nodules (Craig *et al.*, 1999; Gordon *et al.*, 1999) have been isolated to date.

Biochemical and molecular studies of legume nodule metabolism have contributed greatly to our knowledge of primary carbon and nitrogen metabolism in plants, and the interaction between the two (Vance *et al.*, 1994). However, most previous studies focused on one or a few genes/enzymes involved in selected metabolic pathways in many different legume species. We recently published results of a preliminary study of the transcriptome of nodules and roots of the model legume, *Lotus japonicus* (Colebatch *et al.*, 2002). Amongst the 83 nodule-induced genes identified in that study, several were involved in the metabolism of sugars, organic acids, and amino acids. To obtain a broader and deeper view of metabolic differentiation during nodulation, we increased the number of genes represented on cDNA arrays, refined transcriptome data analysis, and combined this with non-biased metabolome analysis. As a result, we identified over 800 nodule-induced genes in *Lotus*, one-third of which are involved in metabolism or transport. We also found significant changes in the nodule metabolome compared with that of roots, which reflected nicely changes in the transcriptome. The results of this work are presented here.

Results

Transcriptome analysis

To facilitate identification of genes involved in SNF in the model legume, *L. japonicus*, we isolated, and sequenced the 5′-end of 100×96 nodule cDNA clones. High-quality sequences were obtained for 8460 clones, and deposited in GenBank with the root name LjNEST (*Lotus japonicus* nodule expressed sequence tag) followed by plate number and alpha-numeric well-coordinates. Approximately 80% of all clones encoded proteins with homologues in public databases. One-quarter of these encoded enzymes or transporters. Most of the enzymes of glycolysis, dicarboxylate synthesis, ammonium assimilation, amino acid biosynthesis, and many more involved in primary and secondary metabolism were represented amongst the sequenced clones. Transporters for metabolites and inorganic ions were also well represented.

A partially redundant DNA array containing 5376 cDNA clones, representing about 2500 genes was constructed and used to compare gene transcript levels in nodules to those in uninfected roots from 7-week-old *Lotus* plants. Partial redundancy amongst clones on the array enabled crosschecking of expression data for many genes. Although a majority of the genes represented on the array were not differentially expressed in nodules compared with roots (Figure 1), approximately 860 genes were induced significantly in nodules (nodule/root ratio >2, $P < 0.05$; Table S1). Over 70% of these had a *P*-value of less than 0.01. Compared with the number of nodule-induced genes discovered, relatively few genes were found to be expressed at a lower level in nodules than in roots (Figure 1), a bias that can be explained by the fact that the arrays were created with nodule cDNA clones.

Approximately one-third of all nodule-induced genes encoded proteins involved in metabolism or transport (28 and 5%, respectively; Figure 2). Eight per cent of nodule-induced genes were predicted to be involved in transcription and its control, while an additional 5% are likely to be involved in signalling. Genes involved in protein synthesis (8%), cell biogenesis (3%), cell division (2%), and intracellular transport processes (2%) were also induced in nodules compared with roots. Approximately one-quarter (26%) of the genes induced in nodules encode proteins that have homologues of unknown function in other species, while the remainder have no known homologues (Figure 2).

Further analysis of the 238 nodule-induced genes encoding enzymes (Table 1), using the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database (http://www.genome.ad.jp/kegg/), showed that many participate in

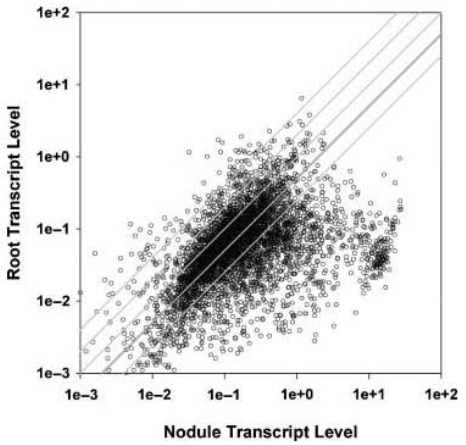


Figure 1. Scatter plot of gene activity (relative transcript level) in *Lotus* nodules and roots.
Comparison of transcript profiles from nodules and roots of 7-week-old *Lotus* plants. Gene activities were determined by a reverse-Northern approach, using cDNA arrays (see Experimental procedures). Points represent the average of eight and five filter hybridizations for nodules and roots, respectively, from two biological replicates in both cases. Lines represent activity ratios of 1 (centre line), 2 and 4.

common metabolic pathways. Thus, many genes involved in starch and sugar metabolism were found to be significantly upregulated in nodules compared with roots. These included two genes encoding starch phosphorylases and two for sucrose synthases, which were induced from four- to ninefold (Table 1). A similar level of induction was observed for genes involved in fructose metabolism, which included a gene for fructokinase, two for pyrophosphate-dependent phosphofructokinase and one for fructose-bisphosphate aldolase. Genes encoding most of the glycolytic enzymes were induced two- to fourfold in nodules compared with roots. These included genes for a cytosolic triose phosphate isomerase, two phosphoglycerate kinases (one cytosolic and one plastidic), two enolases, and a pyruvate kinase.

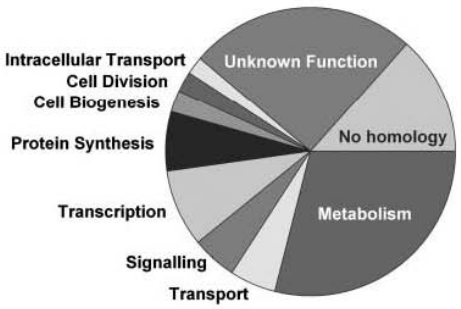


Figure 2. Pie diagram showing functional classes of nodule-induced genes.

Several genes involved in carbon fixation and malate synthesis were found to be upregulated in nodules compared with roots (Table 1). These included two for carbonic anhydrases (CAs) and two for phosphoenolpyruvate carboxylases, which together fix CO_2 into oxaloacetate. Up to seven different genes encoding putative aspartate aminotransferases (AAT) were also nodule-induced (from two to 19-fold), as were genes for two malate dehydrogenases, which convert oxaloacetate to malate, the principal carbon source for bacteroid nitrogen fixation.

A large number of genes involved in amino acid metabolism was induced in mature nodules. Some of these are involved in ammonium assimilation and synthesis of asparagine, the major export form of nitrogen from nodules. These include genes encoding two glutamine synthetases (GS), the seven AAT mentioned above, and two asparagine synthetases (Table 1). Approximately 30 other genes involved in amino acid metabolism were nodule-induced. Amongst these were two genes involved in polyamine biosynthesis (arginine decarboxylase and ornithine decarboxylase) and several genes involved in proline metabolism, including ornithine cyclodeaminase and two encoding proline oxidases.

Genes involved in the synthesis of redox-active compounds such as ascorbate (L-ascorbate peroxidase) and reduced glutathione (glucose-6-phosphate 1-dehydrogenase) were more active in nodules than in roots (Table 1). A gene encoding an oxidoreductase of the 2OG-Fe(II) oxidase family was nearly 30-fold induced during nodulation. Genes encoding a peroxiredoxin and a chloroplastic M-type thioredoxin were also nodule-induced.

Approximately 150 genes involved in other aspects of metabolism were induced in nodules (Table 1). These included genes involved in lipid metabolism, amongst which were genes encoding monogalactosyldiacylglycerol (MGDG) synthase and digalactosyldiacylglycerol (DGDG) synthase and two sulpholipid synthase genes, all of which produce substitutes for phospholipids. Genes for two phosphatidic acid phosphatases, which are involved in phospholipid turnover, and three genes encoding acid phosphatases were also induced in nodules compared with roots. Numerous nodule-induced genes involved in flavonoid and lignin biosynthesis and other secondary metabolism were identified, including genes encoding two 4-coumarate:CoA ligases, caffeoyl-CoA 3-O-methyltransferase, cinnamoyl-CoA reductase, two cinnamyl-alcohol dehydrogenases, flavanone 3-hydroxylase, three dioxygenases, and potentially 13 different cytochromes P-450. Three nodule-induced genes involved in porphyrin metabolism were identified; two encoding plastidic coproporphyrinogen

Table 1 Nodule-induced metabolic genes

EST ID.	Contig*	Ratio (N/R)	*P*	EC#	Annotation
Sucrose and starch					
LjNEST6G11	S	3.7	5.3E-03	2.4.1.1	Starch phosphorylase
LjNEST2D12	TC1815	11.6	1.7E-03	2.4.1.12	Cellulose synthase
LjNEST16B1	TC127*	8.3	1.6E-03	2.4.1.13	Sucrose synthase
LjNEST55G2	TC2075	3.9	8.0E-03	2.4.1.13	Sucrose synthase
LjNEST31B8	S	3.3	2.4E-02	2.4.1.15	Trehalose-6-phosphate synthase
LjNEST39H6	S	2.7	2.6E-02	2.4.1.15	Trehalose-6-phosphate synthase
LjNEST53C6	TC2027*	5.3	6.4E-05	3.1.1.11	Pectinesterase
LjNEST43D4	S	32.6	2.0E-03	3.1.1.11	Pectinesterase
LjNEST45F3	S	2.2	2.4E-02	3.2.1.3	Glucan 1,4-alpha-glucosidase
LjNEST52A1	TC159	3.7	3.1E-03	3.2.1.39	β-1,3-glucanase
LjNEST6C12	TC1455	8.6	2.0E-02		Starch associated protein R1
Fructose and mannose metabolism					
LjNEST25E1	TC2607	4.1	5.0E-04	1.1.1.14	Sorbitol dehydrogenase
LjNEST23B10	TC1980	9.6	3.9E-06	2.7.1.4	Fructokinase
LjNEST45C8	TC2796	6.6	2.1E-03	2.7.1.90	Pyrophosphate-dependent phosphofructokinase
LjNEST2E6	TC1757	4.0	7.0E-03	2.7.1.90	Pyrophosphate-dependent phosphofructokinase
LjNEST3A7	TC687*	4.1	2.5E-02	2.7.7.13	GDP-D-mannose pyrophosphorylase
LjNEST33E7	TC399	5.4	3.7E-03	4.1.2.13	Fructose-bisphosphate aldolase
Glycolysis					
LjNEST31E3	TC2678	3.7	3.4E-03	1.1.1.1	Alcohol dehydrogenase
LjNEST14B1	TC847	4.6	5.1E-03	1.2.1.9	NADP-dep. glyceraldehyde-3-phosphate dehydrogenase
LjNEST38C9	TC329	4.3	2.1E-03	1.2.4.1	Pyruvate dehydrogenase mitochondrial
LjNEST18H9	TC2146	2.7	1.4E-03	2.7.1.40	Pyruvate kinase
LjNEST56G6	TC40	3.7	1.6E-03	2.7.2.3	Phosphoglycerate kinase chloroplastic
LjNEST16A7	TC1983*	2.1	8.6E-03	2.7.2.3	Phosphoglycerate kinase cytosolic
LjNEST39G7	TC1975	2.3	4.4E-02	4.1.1.1	Pyruvate decarboxylase
LjNEST15B6	TC52*	3.2	8.0E-04	4.2.1.11	Enolase
LjNEST43D1	TC53	3.3	2.0E-02	4.2.1.11	Enolase
LjNEST12C12	TC2253	16.9	1.7E-04	5.3.1.1	Triose phosphate isomerase cytosolic
Carbon fixation, dicarboxylate and glyoxylate metabolism					
LjNEST34E11	TC1282	4.6	1.7E-02	1.1.1.29	Glycerate dehydrogenase
LjNEST48F5	S	3.6	1.1E-03	1.1.1.37	Malate dehydrogenase
LjNEST27A12	TC82*	2.7	1.2E-02	1.1.1.37	Malate dehydrogenase
LjNEST25G2	TC2364	4.4	1.2E-05	1.2.1.2	Formate dehydrogenase mitochondrial
LjNEST18H10	S	2.5	1.8E-03	1.5.1.5	Methylenetetrahydrofolate dehydrogenase (NADP+)
LjNEST17G2	TC1769	2.3	4.5E-06	3.5.1.49	Formamidase
LjNEST13G10	TC3075	15.5	1.4E-03	3.5.1.49	Formamidase
LjNEST1H5	S*	2.1	2.9E-05	3.5.1.9	Formamidase
LjNEST16E2	TC373*	2.7	1.1E-02	4.1.1.31	Phosphoenolpyruvate carboxylase
LjNEST3H12	S	3.1	9.1E-03	4.1.1.31	Phosphoenolpyruvate carboxylase
LjNEST20C2	TC1742	2.0	2.9E-02	4.1.1.39	Ribulose bisphosphate carboxylase small chain
LjNEST39B7	TC2283*	8.9	1.2E-05	4.2.1.1	Carbonic anhydrase
LjNEST13E5	S	92.1	1.0E-04	4.2.1.1	Carbonic anhydrase
Glu, Gln, Asp, Asn, Ala					
LjNEST36F9	TC149	7.3	2.0E-04	2.6.1.1	Aspartate aminotransferase
LjNEST6H9	TC1507	5.8	2.3E-04	2.6.1.1	Aspartate aminotransferase
LjNEST2H8	S	19.0	4.4E-04	2.6.1.1	Aspartate aminotransferase
LjNEST20A9	TC542*	6.0	1.2E-03	2.6.1.1	Aspartate aminotransferase
LjNEST53G5	TC150	2.1	4.3E-03	2.6.1.1	Aspartate aminotransferase
LjNEST52A11	TC500	2.8	1.4E-02	2.6.1.1	Aspartate aminotransferase
LjNEST6G3	S	3.0	3.0E-02	2.6.1.1	Aspartate aminotransferase
LjNEST46F1	TC1485	2.8	1.1E-03	2.6.1.2	Alanine aminotransferase
LjNEST12B2	S	2.5	2.6E-02	4.1.1.19	Arginine decarboxylase
LjNEST53D3	TC35*	3.0	7.8E-03	6.3.1.2	Glutamine synthetase
LjNEST11C2	TC549	2.7	1.3E-02	6.3.1.2	Glutamine synthetase-like
LjNEST3A4	TC91*	4.8	9.0E-04	6.3.5.4	Asparagine synthase
LjNEST17E9	S	Inf	1.9E-02	6.3.5.4	Asparagine synthase

Table 1 continued

EST ID.	Contig*	Ratio (N/R)	*P*	EC#	Annotation
Other amino acids					
LjNEST2A6	S	2.7	1.6E-03	1.1.1.35	Acyl-CoA synthetase
LjNEST15F6	TC2381	4.9	1.0E-03	1.1.1.85	Isicitrate dehydrogenase
LjNEST23G3	TC798	2.0	9.8E-03	1.1.1.95	Phosphoglycerate dehydrogenase
LjNEST47H8	TC2516	6.0	1.2E-03	1.11.1.7	Peroxidase
LjNEST17H9	TC2005	9.2	9.9E-03	1.11.1.7	Peroxidase
LjNEST19F3	S	5.6	5.9E-04	1.13.11.5	Homogentisate 1,2-dioxygenase
LjNEST43C8	TC2959	5.2	5.1E-05	1.14.11.2	Prolyl 4-hydroxylase
LjNEST45D5	TC2947	8.2	2.5E-03	1.14.11.2	Procollagen-proline dioxygenase
LjNEST14A10	S	4.8	1.1E-04	1.2.1.8	Betaine aldehyde dehydrogenase
LjNEST6A8	S	3.3	1.9E-02	1.4.3.6	Copper amine oxidase
LjNEST15E6	TC1946	4.4	1.1E-03	1.5.3.-	Proline oxidase
LjNEST11E4	TC635	2.5	1.3E-03	1.5.99.8	Proline oxidase mitochondrial
LjNEST16E1	TC877	3.3	9.2E-03	2.1.1.-	Protein arginine N-methyltransferase
LjNEST53C12	TC539	4.4	3.3E-03	2.1.1.14	Methionine synthase
LjNEST6H6	S	10.4	1.6E-04	2.6.1.42	Branched-chain amino acid aminotransferase
LjNEST8H5	TC2989	4.9	5.7E-03	2.6.1.42	Branched-chain amino acid aminotransferase
LjNEST55H10	TC2504	3.8	3.5E-03	2.7.1.39	Homoserine kinase
LjNEST50D5	TC502	2.0	3.9E-02	3.1.2.4	3-hydroxyisobutyryl-coenzyme A hydrolase
LjNEST42C3	S	3.1	1.5E-03	3.2.1.147	Myrosinase
LjNEST10B2	TC3618	2.3	3.8E-02	3.2.2.9	S-adenosyl homocysteine nucleosidase
LjNEST35A2	TC3140	7.2	4.2E-04	3.5.1.-	Amidase
LjNEST42A8	S	3.4	6.0E-04	3.5.1.-	N-carbamyl-L-amino acid amidohydrolase
LjNEST10B11	TC1947*	7.7	3.7E-04	3.5.5.1	Nitrilase
LjNEST37B10	TC496	7.2	3.5E-04	4.1.1.17	Ornithine decarboxylase
LjNEST54C2	TC374	6.6	1.8E-03	4.1.3.12	2-isopropylmalate synthase
LjNEST29A1	S	28.2	3.4E-03	4.1.3.12	2-isopropylmalate synthase
LjNEST2A5	S	13.1	1.2E-03	4.1.3.12	2-isopropylmalate synthase
LjNEST5B1	S	13.7	1.6E-02	4.1.3.27	Anthranilate synthase
LjNEST27E11	TC171	6.6	2.1E-07	4.2.99.8	Cysteine synthase
LjNEST6C9	S	3.8	1.1E-02	4.2.99.8	Cysteine synthase
LjNEST48F3	TC2718	2.8	2.9E-03	4.3.1.12	Ornithine cyclodeaminase
LjNEST18H8	TC1187	2.8	4.3E-03	4.3.1.5	Phenylalanine ammonia-lyase 1
LjNEST29B1	TC2226*	14.2	7.7E-05	5.4.99.5	Chorismate mutase
LjNEST37F9	S	20.8	2.0E-03	5.4.99.5	Chorismate mutase
Purine metabolism					
LjNEST22C12	TC2544	10.3	3.9E-06	1.1.1.205	Inosine-5′-monophosphate dehydrogenase
LjNEST51A10	TC2150	3.7	3.8E-04	1.7.3.3	Uricase
LjNEST13H5	TC3335	13.2	1.4E-06	2.1.2.3	IMP cyclohydrolase
LjNEST27D6	TC131	28.5	2.4E-03	2.4.2.7	Adenine phosphoribosyltransferase
LjNEST53D10	TC1115	7.2	4.0E-03	2.7.4.6	Nucleoside diphosphate kinase
LjNEST8E7	TC3156	2.1	1.4E-02	2.7.6.5	GTP pyrophosphokinase
Redox metabolism					
LjNEST23D11	S	7.9	1.2E-03	1.1.1.49	Glucose-6-phosphate 1-dehydrogenase
LjNEST45C9	TC779	2.2	4.9E-03	1.11.1.11	L-ascorbate peroxidase
LjNEST39G1	TC321	3.4	8.7E-03	1.6.4.-	Peroxiredoxin
LjNEST6F9	TC1132	51.3	7.9E-03		Thioredoxin M-type chloroplastic
LjNEST6H2	S	29.4	3.5E-05		Oxidoreductase of 2OG-Fe(II) oxidase family
Other metabolism					
LjNEST34D8	TC2479	5.1	1.6E-03	1.-.-.-	Aldo/keto reductase
LjNEST51C4	TC786	5.6	1.8E-02	1.1.1.100	Short chain alcohol dehydrogenase
LjNEST47H3	S	3.5	2.0E-03	1.1.1.100	Short chain alcohol dehydrogenase
LjNEST19D1	TC616	Inf	1.2E-04	1.1.1.195	Cinnamyl-alcohol dehydrogenase
LjNEST23F4	S	14.4	2.2E-03	1.1.1.195	Cinnamyl-alcohol dehydrogenase
LjNEST29G12	S	23.1	1.7E-02	1.1.1.51	Hydroxysteroid (17-β) dehydrogenase
LjNEST51E8	TC2034	3.9	4.8E-03	1.10.2.2	Ubiquinol-cytochrome C reductase mitochondrial
LjNEST13A6	S	6.8	3.8E-05	1.13.11.-	Dioxygenase
LjNEST41C12	TC647	4.6	2.8E-03	1.13.11.-	Dioxygenase

Table 1 continued

EST ID.	Contig*	Ratio (N/R)	*P*	EC#	Annotation
LjNEST49B5	TC2679	3.3	9.5E-04	1.13.11.-	Dioxygenase
LjNEST5H5	TC108	4.8	2.9E-04	1.13.11.12	Lipoxygenase
LjNEST17D7	S	26.4	2.7E-03	1.14.-.-	Cytochrome P450
LjNEST55F10	TC3403	21.5	1.8E-03	1.14.-.-	Cytochrome P450
LjNEST14A8	S	6.8	2.0E-05	1.14.-.-	Cytochrome P450
LjNEST18A1	S	4.4	1.0E-04	1.14.-.-	Cytochrome P450
LjNEST4A12	S	4.3	2.7E-02	1.14.-.-	Cytochrome P450
LjNEST38G6	TC2301*	4.0	3.4E-04	1.14.-.-	Cytochrome P450
LjNEST49C3	S	3.9	3.3E-04	1.14.-.-	Cytochrome P450
LjNEST24D4	TC1231	3.4	1.8E-03	1.14.-.-	Cytochrome P450
LjNEST48D3	TC2914	3.1	4.4E-03	1.14.-.-	Fatty acid hydroperoxide lyase
LjNEST15F4	TC2706	3.0	3.4E-04	1.14.-.-	Cytochrome P450
LjNEST2H12	TC408	2.7	4.2E-02	1.14.-.-	Cytochrome P450
LjNEST6A2	S	2.3	5.7E-03	1.14.-.-	Cytochrome P450
LjNEST48C2	S	2.2	3.2E-02	1.14.-.-	Cytochrome P450
LjNEST41E9	TC3717	2.0	4.8E-02	1.14.-.-	Cytochrome P450
LjNEST8H10	TC75	4.1	3.2E-04	1.14.1.-	Alpha-dioxygenase
LjNEST15A10	TC2382*	9.8	2.8E-04	1.14.11.-	Flavanone 3-hydroxylase
LjNEST33C6	S	2.7	1.5E-03	1.14.13.72	C-4 sterol methyl oxidase
LjNEST15D6	TC3204	9.3	6.3E-04	1.14.13.8	Flavin-containing monooxygenase
LjNEST11H7	TC2208	4.5	1.3E-04	1.14.14.1	Obtusifoliol 14-demethylase
LjNEST49F6	TC2057	2.4	7.2E-03	1.14.14.1	Obtusifoliol 14-demethylase
LjNEST55H3	TC455	3.8	1.7E-02	1.14.99.-	Delta8 sphingolipid desaturase
LjNEST4C2	TC1374	4.2	6.4E-04	1.14.99.3	Haem oxygenase
LjNEST24E11	TC3759	7.9	6.5E-03	1.14.99.6	Stearoyl acyl carrier protein desaturase
LjNEST25G4	TC2599	4.8	1.2E-05	1.2.1.44	Cinnamoyl-CoA reductase
LjNEST42B5	S	2.2	1.9E-02	1.3.1.22	Steroid 5alpha-reductase
LjNEST56E9	S	2.9	4.2E-03	1.3.1.45	2'-hydroxyisoflavone reductase
LjNEST24B9	TC354*	94.9	5.0E-05	1.3.3.3	Coproporphyrinogen III oxidase chloroplastic
LjNEST24G3	TC1785	10.8	4.9E-08	1.3.3.3	Coproporphyrinogen III oxidase chloroplastic
LjNEST44D5	TC1278	3.2	4.6E-03	1.3.3.6	Acyl-CoA oxidase
LjNEST49E1	TC2493	3.4	7.6E-03	1.3.99.1	Succinate dehydrogenase
LJNEST28F7	TC223	2.5	6.8E-04	1.4.3.-	1-aminocyclopropane-1-carboxylate oxidase
LjNEST8A9	TC1150	5.1	6.9E-04	1.6.2.2	NADH-cytochrome b5 reductase
LjNEST15A1	S	2.0	2.9E-03	1.6.5.3	NADH:ubiquinone oxidoreductase mitochondrial
LjNEST52E2	S	7.1	2.6E-02	1.6.5.5	Quinone oxidoreductase
LjNEST15A11	TC2173*	17.3	9.0E-05	2.1.1.129	O-methyltransferase
LjNEST1D6	S	2.3	1.4E-02	2.1.1.104	Caffeoyl-CoA 3-O-methyltransferase
LjNEST45G9	TC105	2.7	6.7E-03	2.1.1.41	Sterol-C-methyltransferase *Arabidopsis thaliana*
LjNEST33F3	TC2246	8.0	2.5E-03	2.3.1.110	Tyramine hydroxycinnamoyltransferase
LjNEST34D9	TC3413	4.1	7.7E-03	2.3.1.47	8-amino-7-oxononanoate synthase
LjNEST28F8	S	5.0	3.8E-02	2.4.-.-	Glycosyl transferase
LjNEST40C10	TC2191	3.0	7.4E-04	2.4.-.-	Glycosyl transferase
LjNEST29E4	S	2.3	3.3E-02	2.4.-.-	Glycosyl hydrolase
LjNEST44D11	TC3427	58.8	6.6E-06	2.4.1.-	UDP-glucosyltransferase
LjNEST4B4	TC1641	38.2	1.8E-03	2.4.1.-	Xyloglucan fucosyltransferase
LjNEST23E1	TC286	23.6	1.1E-04	2.4.1.-	Glucosyltransferase
LjNEST31F8	TC1584	5.1	4.0E-02	2.4.1.-	Glycosyltransferase
LjNEST30B6	TC451*	9.6	3.3E-02	2.4.1.121	Indole-3-acetate β-glucosyltransferase
LjNEST17A3	TC853	4.4	1.6E-03	2.4.1.46	MGDG synthase
LjNEST12D12	TC1344*	3.5	5.3E-03	2.5.1.-	Ent-kaurene synthase A
LjNEST36D9	TC2467	5.9	2.8E-03	2.5.1.18	Glutathione S-transferase
LjNEST23H1	TC750	3.1	2.1E-02	2.5.1.18	Glutathione S-transferase
LjNEST38B8	S	15.5	2.8E-04	2.5.1.29	Geranylgeranyl pyrophosphate synthase
LjNEST12B5	S	6.8	1.6E-03	2.5.1.29	Geranylgeranyl pyrophosphate synthase
LjNEST1H10	S	14.3	1.3E-03	2.7.7.-	Phosphoglyceride transfer protein
LjNEST11G12	TC3086	2.1	1.1E-02	2.7.7.15	Choline-phosphate cytidylyltransferase
LjNEST14A12	TC2721	4.2	3.0E-04	2.7.7.23	UDP-N-acetylglucosamine pyrophosphorylase
LjNEST52D1	S	3.6	4.1E-03	2.7.8.23	Phosphoenolpyruvate mutase
LjNEST27D4	S	6.1	9.1E-03	3.1.1.-	Pectin acetylesterase

Table 1 continued

EST ID.	Contig*	Ratio (N/R)	*P*	EC#	Annotation
LjNEST35D1	S	18.8	6.6E-04	3.1.1.-	Lipase/hydrolase
LjNEST19F4	S	8.4	4.4E-02	3.1.1.-	Lipase
LjNEST13H2	TC2709	11.3	8.3E-05	3.1.1.1	Esterase
LjNEST29D5	S	7.3	3.3E-04	3.1.1.1	Esterase
LjNEST52F3	TC834	4.9	4.6E-04	3.1.3.2	Acid phosphatase
LjNEST47C3	TC2335	3.7	6.9E-05	3.1.3.2	Acid phosphatase-PAP
LjNEST4A10	TC866	2.1	9.3E-04	3.1.3.2	Acid phosphatase
LjNEST47E7	S	3.9	1.6E-03	3.1.3.4	Phosphatidic acid phosphatase
LjNEST27F6	S	7.8	2.7E-03	3.1.3.4	Phosphatidic acid phosphatase
LjNEST48B3	TC2113	3.5	2.1E-02	3.2.1.14	Chitinase
LjNEST54G8	TC1675	4.6	1.7E-02	3.2.1.22	Alpha galactosidase
LjNEST17H5	TC2753	5.4	9.0E-06	3.2.1.51	Profucosidase
LjNEST48E7	TC1806	4.1	2.0E-02	3.2.1.51	Alpha-fucosidase
LjNEST54B7	TC236	3.2	3.0E-04	3.3.2.3	Epoxide hydrolase
LjNEST2B8	TC235	2.1	1.3E-02	3.3.2.3	Epoxide hydrolase
LjNEST35F12	TC2104	8.3	1.8E-03	3.4.19.9	Gamma glutamyl hydrolase
LjNEST52A9	S	18.5	6.4E-04	4.1.2.25	Dihydroneopterin aldolase
LjNEST1H7	TC2436	6.5	9.4E-05	4.1.2.25	Dihydroneopterin aldolase
LjNEST9E4	TC759	9.2	9.5E-05	4.1.3.8	ATP citrate lyase b-subunit
LjNEST4A5	S	3.5	3.9E-02	4.1.3.8	ATP-citrate lyase
LjNEST52F8	TC789	2.6	4.7E-03	4.2.1.46	dTDP-glucose 4-6-dehydratase
LjNEST15B7	TC24*	9.0	2.5E-05	4.2.1.52	Thiamin biosynthetic enzyme
LjNEST45G11	TC1411	5.2	4.2E-04	4.2.2.2	Pectate lyase
LjNEST23G7	TC782	2.2	6.3E-03	4.4.1.18	Prenylcysteine lyase
LjNEST3A5	TC3774	46.7	8.6E-04	5.1.3.2	Epimerase/dehydratase
LjNEST29F4	TC1536	5.9	2.5E-02	5.1.3.2	Epimerase
LjNEST46G5	TC504*	4.9	1.7E-03	5.1.3.2	Epimerase/dehydratase
LjNEST42C1	TC753	2.9	1.3E-03	5.2.1.8	Peptidyl-prolyl cis/trans isomerase
LjNEST35A1	TC2418	2.3	1.4E-03	5.2.1.8	Peptidylprolyl isomerase
LjNEST49G9	S	9.4	5.1E-06	5.5.-.-	Cycloisomerase
LjNEST3C2	TC2406	2.8	4.1E-05	5.5.1.4	Myo-inositol-1-phosphate synthase
LjNEST18F10	TC2692	14.8	2.0E-02	6.2.1.12	4-coumarate:CoA ligase
LjNEST18A2	TC3751	3.0	1.9E-02	6.2.1.12	4-coumarate–CoA ligase
LjNEST14C2	S	3.1	4.9E-02	6.3.2.17	Folylpolyglutamate synthetase
LjNEST54B1	TC661	2.4	3.4E-03	6.3.4.-	Biotin holocarboxylase synthetase
LjNEST55C4	S	3.6	3.4E-03	6.4.1.2	Acetyl-CoA carboxylase
LjNEST27D7	TC1581	40.2	1.4E-03		10-deacetylbaccatin III-10-O-acetyl transferase
LjNEST45D3	TC2687	11.0	1.0E-05		Acetone-cyanohydrin lyase
LjNEST55B8	TC94	2.4	4.6E-03		ADP ribosylation factor
LjNEST37G3	TC2102	2.5	4.0E-03		ADP-ribosylation factor
LjNEST6H1	S	6.7	6.2E-05		Anthocyanin 5-aromatic acyltransferase
LjNEST46B10	S	5.1	2.9E-03		Anthocyanin 5-aromatic acyltransferase
LjNEST28E1	S	4.4	1.9E-07		Anthranilate N-hydroxycinnamoyl/benzoyltransferase
LjNEST17F11	S	2.6	9.5E-03		β-amyrin synthase
LjNEST4F1	TC1489	10.8	2.1E-02		β-ketoacyl-CoA synthase
LjNEST26A10	TC975	15.0	4.9E-07		Chalcone reductase
LjNEST25A5	S	6.5	4.9E-04		Cytochome b5
LjNEST44D2	TC2614	3.0	7.2E-03		Cytochrome c
LjNEST48E6	TC1482	5.1	2.2E-04		Cytokinin synthase
LjNEST15E4	S	3.9	5.2E-04		Digalactosyldiacylglycerol synthase
LjNEST25D4	S	3.2	4.5E-02		Fatty acid elongase 3-ketoacyl-CoA synthase
LjNEST47B10	S	2.2	2.4E-02		Fatty acid elongase 3-ketoacyl-CoA synthase
LjNEST41G5	TC636	2.9	2.2E-03		Flavonol glucosyltransferase
LjNEST11E10	S	23.6	3.9E-02		Hydrolase
LjNEST5A6	TC1306	9.0	2.8E-02		Hydrolase
LjNEST36A2	TC2702	2.8	5.7E-03		Hydrolase
LjNEST46A4	TC2591	4.3	1.3E-04		Hydrolase
LjNEST5B11	TC1813	156.4	9.4E-04		Lipoyltransferase
LjNEST42A12	S	2.2	1.6E-02		Mitochondrial uncoupling protein
LjNEST15B6	TC2110	3.6	1.7E-03		Myo-inositol oxygenase

Table 1 continued

EST ID.	Contig*	Ratio (N/R)	*P*	EC#	Annotation
LjNEST2B9	S	9.0	3.0E-03		N-hydroxycinnamoyl/benzoyltransferase
LjNEST34F12	S	3.8	2.8E-03		N-hydroxycinnamoyl/benzoyltransferase
LjNEST42E5	S	2.4	1.3E-02		NifU-like metallocluster assembly factor
LjNEST3H2	S	7.5	2.7E-05		Nucleotide-binding protein
LjNEST11G8	TC3186*	18.5	8.0E-07		Phosphatidylinositol transfer protein
LjNEST3G12	TC1093*	648.7	2.3E-05		Phytocyanin
LjNEST46F10	TC146	2.0	3.1E-02		Plastocyanin a
LjNEST46B5	TC11	5.9	3.9E-03		Rubisco activase
LjNEST2D4	S	4.6	5.6E-03		Sulpholipid synthase
LjNEST46F8	TC3002	2.8	5.2E-03		Sulpholipid synthase
LjNEST40C11	S	5.3	1.1E-02		Terminal oxidase plastidic
LjNEST51B8	S	10.5	3.3E-04		Ubiquinone biosynthesis protein

EST identifier and corresponding tentative consensus number from the TIGR *Lotus japonicus* Gene index (http://www.tigr.org/tdb/tgi/ljgi/) are shown. Corresponding GenBank accession numbers are provided in Table S1. N/R denotes mean nodule/root transcript ratio. *P*-values were obtained from Student's *t*-tests. Enzyme commission numbers and automatic annotations are also shown. *Previously found to be induced in nodules (Colebatch *et al.*, 2002).

III oxidase, involved in haem biosynthesis, and one encoding haem oxygenase, which is involved in haem degradation. Transcripts of the coproporphyrinogen III oxidase genes were between 10 and 90 times more abundant in nodules than in roots.

Genes involved in plant hormone metabolism were also upregulated in nodules compared with roots: one encoding ent-kaurene synthase A and one for cytokinin synthase are involved in synthesis of gibberellic acid (GA) and cytokinins, respectively, while the gene for indole-3-acetate (IAA) β-glucosyltransferase directs inactivation of IAA.

Metabolic differentiation during nodule development was reflected at another level. Almost 50 genes encoding transporters were upregulated in nodules compared with roots (Table 2). These included genes for four related putative sugar transporters. Genes encoding four putative peptide transporters and a related low-affinity nitrate transporter were highly upregulated in nodules, as were genes for four sulphate transporter homologues. Genes encoding two porins, two ABC transporters, and a variety of ATPases were also induced in nodules, as were two Na^+/H^+ antiporter genes.

Little is known about signalling or control of transcription in mature, nitrogen-fixing root nodules. We identified 113 genes in these two categories that were more than twofold upregulated in nodules compared with roots (Table 3). Of the 43 genes potentially involved in signalling processes, 14 encoded putative kinases including three receptor kinases. Seven phosphatase genes were also identified. Homologues of a variety of genes involved in plant–microbe interactions were identified: an NBS-LRR and an LRR gene, two *Mlo* genes, and genes encoding two other disease resistance proteins. Putative ethylene receptor and response regulator genes were also induced in nodules (Table 3).

A large number of novel transcription factor (TF) genes were found to be nodule-induced, amongst them genes encoding four homeodomain proteins, three bZIP TFs, three MYB family TFs, and 10 zinc finger proteins. Many other genes potentially involved in transcription or RNA processing were also induced in nodules (Table 3). TF genes typically are expressed at very low levels (Czechowski *et al.*, 2004), and this was the case in the experiments described here. Real-time RT-PCR was performed on all (20) TF genes in Table 3 with relative root transcript levels less than 0.2. Transcript levels for slightly more than half of these genes (11/20) were found to be significantly higher in nodules than in roots, using this method (Table 4). In four cases, TC902, TC3645, LjNEST55C5, and TC367, the nodule/root transcript ratios obtained from RT-PCR were much greater than those obtained from the cDNA array. None of these genes were identified as nodule-induced when the method of Colebatch *et al.* (2002) was used to analyse the current array data. Therefore, by refining the earlier method (see Experimental procedures) we were able to uncover potentially important nodule-induced regulatory genes. However, this came at the cost of an increased rate of false positives amongst genes expressed at low levels. In contrast, false-positive results for genes expressed at higher levels in nodules were comparatively rare (relative transcript level >0.2; nodule/root ratio >2; Table 4). All seven genes in this category were confirmed as nodule-induced by real-time RT-PCR (Table 4).

Genes involved in protein synthesis, processing, and turnover represented 8% of all nodule-induced genes (Table S2). These included numerous genes for ribosomal subunits, translation initiation and elongation, chaperonins, and various peptidases and proteases. Nodule-induced genes potentially involved in controlled protein degradation encode various proteasome subunits, the COP9 complex

Table 2 Nodule-induced transporter genes

EST ID.	Contig*	Ratio (N/R)	*P*	Annotation
LjNEST2B12	TC721	297.8	1.6E-07	Peptide transporter
LjNEST7E5	TC3742	174.6	1.5E-03	Na+/H+ antiporter
LjNEST37E2	TC547*	146.0	5.2E-04	Sulphate transporter
LjNEST1G2	TC1539*	81.1	1.5E-04	Membrane protein
LjNEST12E11	TC2128	65.4	3.1E-04	Sulphate transporter
LjNEST23B9	TC662	47.5	4.8E-07	Peptide transporter
LjNEST24C7	TC3063	37.5	8.2E-05	Metal transporter Nramp family
LjNEST18E3	S	24.6	1.5E-03	Sodium/hydrogen exchanger
LjNEST15C6	TC1236	19.4	2.9E-04	Low-affinity nitrate transporter (PTR/POT family)
LjNEST48F9	S	13.0	9.1E-05	Sugar transporter
LjNEST16C11	TC1688	12.8	1.4E-04	Membrane protein, MtN21-like
LjNEST46B6	TC397	10.6	9.3E-04	Sugar transporter
LjNEST5B12	S	9.3	2.0E-03	Peptide transporter
LjNEST22E8	TC2539	9.3	2.9E-04	Oligopeptide transporter
LjNEST19B11	TC398	8.6	3.1E-04	Sugar transporter
LjNEST20F7	TC2686	8.4	4.5E-02	Sulphate transporter
LjNEST52E4	S	7.3	3.2E-02	Plastid inner envelope membrane protein
LjNEST4A6	S	6.8	3.4E-02	Membrane protein
LjNEST53B9	S	6.6	6.3E-06	Peroxisomal membrane protein
LjNEST5D1	S	6.2	1.5E-04	Plastid inner envelope membrane protein
LjNEST45D6	TC2999	6.2	7.7E-03	Purine permease
LjNEST30G7	TC823	6.1	2.4E-03	ABC transporter
LjNEST19C8	TC510	5.2	5.5E-05	V-ATPase A subunit
LjNEST55F12	S	4.6	4.6E-03	Sulphate transporter
LjNEST15F7	S	4.4	3.4E-02	V-ATPase D subunit
LjNEST5E1	TC804	4.3	6.7E-05	ATP synthase epsilon chain
LjNEST44E3	TC765	4.2	2.5E-02	Phosphate transporter
LjNEST50F3	S	4.1	3.4E-03	Peroxisomal membrane protein
LjNEST44A4	S	3.9	4.8E-03	Mitochondrial inner membrane protein
LjNEST54A1		3.8	4.7E-04	Membrane protein
LjNEST49D12	TC2071	3.7	2.5E-04	Membrane protein
LjNEST54D1	TC740	3.7	3.8E-03	Membrane protein
LjNEST14B12	S	3.6	3.5E-04	Phosphate/phosphoenolpyruvate translocator
LjNEST5E8	TC2428	3.3	6.1E-03	Porin
LjNEST55C3	S	3.2	6.9E-03	Multidrug-resistance related protein, ABC transporter
LjNEST54A3	S	2.9	1.2E-03	Histidine transporter
LjNEST45H6	S	2.9	1.0E-02	Cationic amino acid transporter
LjNEST37G10	TC1928	2.9	4.5E-02	Triose phosphate/phosphate translocator, chloroplastic
LjNEST28A11	TC1993	2.8	2.7E-04	Porin
LjNEST27A2	TC2010	2.6	1.6E-05	LjN70-like
LjNEST45A12	TC64	2.5	1.5E-03	ATP synthase gamma chain mitochondrial
LjNEST51G11	TC240	2.5	3.3E-02	Type 1 membrane protein
LjNEST47B2	TC3408	2.4	4.5E-03	V-ATPase A subunit
LjNEST29F8	S	2.4	3.5E-02	Membrane protein
LjNEST37G5	TC301	2.2	1.3E-02	H+-transporting ATP synthase chain 9
LjNEST11G3	TC3638	2.0	1.0E-02	Plasma membrane intrinsic protein, PIP

See Table S2 for an explanation of abbreviations.
*Previously found to be induced in nodules (Colebatch *et al.*, 2002).

subunit CSN2, two F-box proteins, three RING-H2 zinc finger proteins, ubiquitin activating enzyme E1, two ubiquitin conjugating enzymes, and a ubiquitin carboxyl terminal hydrolase.

Relatively few genes involved in cell biogenesis, cell division, and intracellular transport, were expressed at higher levels in mature nodules than in roots (Figure 2; Table S2). Notable exceptions include genes for five different leghaemoglobins, which were the most highly expressed of all genes in nodules, and several genes for small GTP-binding proteins that are presumably involved in vesicle trafficking.

Relatively few genes were found to be repressed in nodules compared with roots (Table S3), as expected, given the nodule source of clones on the arrays. These included genes encoding three putative aquaporins: two PIP homologues and one TIP homologue, which were repressed between two and fivefold in nodules. Transcripts of several genes often associated with plant defence were also signi-

Table 3 Nodule-induced genes for signalling and transcription

EST ID.	Contig*	Ratio (N/R)	*P*	Annotation
Signalling				
LjNEST55C8	S	3.2	4.5E-04	14-3-3 protein
LjNEST40G10	S	3.1	4.0E-02	Ankyrin-repeat protein
LjNEST13H7	TC991	2.4	6.1E-04	Annexin
LjNEST42D5	TC3570	5.1	2.0E-04	ATP/GTP nucleotide-binding protein
LjNEST16F11	S	4.5	1.9E-02	B' regulatory subunit of PP2A
LjNEST11G7	TC1202	2.3	2.0E-03	Calcium-binding protein
LjNEST34C9	S	81.5	6.2E-03	Calcium-dependent protein kinase-like protein
LjNEST30F5	TC100*	6.2	3.5E-06	Calmodulin
LjNEST56F9	TC278*	4.8	9.9E-05	Calmodulin
LjNEST4D2	TC1410	4.1	2.2E-02	Disease resistance protein homolog
LjNEST47D6	TC1266	6.1	1.4E-04	Ethylene receptor (ETR5)
LjNEST11G1	TC3053	4.4	1.6E-03	Glucokinase-associated phosphatase
LjNEST32A8	TC2371	3.7	5.8E-03	GTP-binding protein
LjNEST48H4	TC1428	2.5	3.8E-02	Iron-deficiency specific, Ids4-like protein
LjNEST26H6	TC474	3.5	1.9E-04	Kelch repeat containing F-box protein family
LjNEST27A7	TC2161	10.2	3.9E-04	Leucine-rich repeat protein
LjNEST48G9	TC2786	2.6	5.0E-04	MAP kinase
LjNEST11D10	S	6.9	3.3E-04	Mlo protein
LjNEST4H6	S	10.2	5.3E-06	Mlo protein
LjNEST18E9	TC1535	3.0	4.5E-03	NBS-LRR type protein
LjNEST47H1	S	3.0	1.4E-02	Non-race specific disease resistance protein
LjNEST27B6	TC289	11.2	1.9E-03	Phosphoenolpyruvate carboxylase kinase
LjNEST40A8	S	4.8	4.0E-02	Protein kinase
LjNEST9B3	S	2.9	1.2E-03	Protein kinase
LjNEST6A6	S	2.6	3.1E-02	Protein phosphatase
LjNEST52E9	S	2.2	8.3E-03	Protein phosphatase
LjNEST7H9	TC806	32.6	2.1E-03	Protein phosphatase type 2A
LjNEST38B4	S	6.2	4.8E-04	Protein phosphatase type 2C
LjNEST49E3	S	4.9	4.0E-02	Protein phosphatase type 2C
LjNEST15E5	TC1402	4.6	2.7E-04	Protein phosphatase type 2C
LjNEST16G1	S	2.4	4.0E-04	Protein phosphatase type 2C
LjNEST27G2	TC1433	11.4	5.2E-06	Receptor kinase
LjNEST55B7	S	2.2	1.2E-02	Receptor kinase
LjNEST51F11	TC2164	18.0	5.6E-03	Receptor kinase common family
LjNEST51G2	S	3.8	1.2E-02	Receptor-like protein
LjNEST6D3	S	21.6	2.5E-04	Remorin
LjNEST45C7	TC1832	4.8	3.5E-03	Response regulator
LjNEST56G5	S	3.1	1.1E-03	Serine/threonine protein kinase
LjNEST20A4	TC261	2.5	3.5E-02	Serine/threonine protein kinase
LjNEST46C3	TC1614	2.2	8.5E-03	Serine/threonine protein kinase
LjNEST44B3	TC3045	8.2	5.4E-04	Shaggy-like protein kinase
LjNEST10A10	TC3754	4.5	2.8E-03	SNF1-like protein kinase
LjNEST29G4	TC1102	32.5	9.4E-05	SOS2-like protein kinase
Transcription				
LjNEST36E6	TC2291	2.3	9.6E-03	Arginine/serine-rich splicing factor
LjNEST50C12	TC1753	4.0	1.0E-04	ATP-dependent RNA helicase
LjNEST3H6	TC870	25.0	8.8E-05	BEL1-like homeodomain protein
LjNEST28A12	S	10.2	7.3E-04	BEL1-like homeodomain protein
LjNEST6C7	TC349	49.2	8.0E-05	bZIP transcription factor
LjNEST15F10	TC922	3.6	9.5E-05	bZIP transcription factor
LjNEST6F1	TC1391	3.4	3.0E-02	bZIP transcription factor
LjNEST6A3	TC902	3.6	5.6E-04	CCAAT-box-binding transcription factor
LjNEST52F1	TC2729	3.1	5.4E-03	CCR4-associated transcription factor
LjNEST14H12	TC3440*	17.7	1.9E-04	Chloroplast nucleoid DNA binding protein
LjNEST17A9	TC981	7.3	1.8E-03	DNA-binding protein
LjNEST12E3	TC2417	4.9	4.3E-04	DNA-binding protein
LjNEST50D8	TC881	3.2	2.2E-02	DNA-binding protein
LjNEST44C3	TC2403	3.0	1.8E-03	DNA-binding protein
LjNEST15B4	TC3154	2.3	1.2E-03	DNA-binding protein

Table 3 continued

EST ID.	Contig*	Ratio (N/R)	*P*	Annotation
LjNEST37E4	TC1634	2.9	1.2E-02	DNA-directed RNA polymerase chain III
LjNEST43F1	TC751	2.6	6.4E-03	EREBP-like
LjNEST1A2	TC3037	24.4	1.0E-04	G-box binding protein
LjNEST10E7	TC3007	2.3	3.1E-02	Homeobox RRM-containing protein
LjNEST18G12	TC2543	7.6	9.8E-04	Homeodomain transcription factor
LjNEST36G7	TC1895	5.9	1.6E-06	KH domain/zinc finger protein
LjNEST36E10	TC326	2.6	1.4E-02	KH domain/zinc finger protein
LjNEST14B3	TC1629	2.4	1.4E-05	KH domain/zinc finger protein
LjNEST46H8	TC3654*	4.2	3.3E-02	MADS-box protein
LjNEST40D9	TC367	11.6	2.1E-05	MYB family transcription factor
LjNEST16G12	TC1091	2.3	2.1E-02	MYB family transcription factor
LjNEST11A2	S	2.2	9.2E-03	MYB family transcription factor
LjNEST49H11	TC2380	5.0	1.5E-04	Negative regulator of URS2 of the HO promoter
LjNEST56G8	TC3549	4.5	2.2E-02	Non-LTR retroelement reverse transcriptase
LjNEST10E11	TC543	4.2	4.7E-04	NTF2-containing RNA-binding protein
LjNEST56C5	S	3.4	1.4E-02	PHD-type zinc finger protein
LjNEST22F4	S	10.6	3.6E-04	RNA helicase
LjNEST37C8	TC755	2.9	3.9E-03	RNA helicase
LjNEST18E11	S	2.8	1.8E-02	RNA polymerase II subunit
LjNEST27B3	TC2716	9.0	1.3E-02	RNA polymerase II transcriptional regulation mediator
LjNEST49G12	TC3747	8.2	9.4E-03	RNA-binding protein
LjNEST11F9	TC2958	4.0	2.2E-02	RNA-binding protein
LjNEST54F12	TC2072	2.4	7.1E-03	RNA-binding protein
LjNEST9E10	TC2373	2.3	2.3E-02	RNA-binding protein
LjNEST42B9	TC2710	2.2	5.9E-03	RNA-binding protein
LjNEST16E9	TC3032	2.1	4.2E-02	RNA-binding protein
LjNEST17B4	S	2.0	4.9E-02	RNA-binding protein
LjNEST28C7	TC1892	31.1	4.6E-06	RNA-binding protein
LjNEST12F2	TC89	3.5	4.7E-03	RNA-binding protein
LjNEST27B2	S	27.4	1.2E-02	RNA-binding Sun protein-like
LjNEST12D5	S	20.4	1.8E-03	RNase H
LjNEST37E8	S	2.3	1.1E-02	RNase H
LjNEST15C8	TC1322	22.2	1.2E-04	Splicing factor
LjNEST17D3	TC3424	8.4	1.5E-05	Splicing factor
LjNEST29F2	TC997	7.0	3.5E-03	Splicing factor
LjNEST20H7	TC3702	6.5	1.2E-04	Splicing factor
LjNEST38H1	S	2.2	3.6E-03	Splicing factor
LjNEST40A10	S	2.1	1.6E-02	Squamosa promoter binding protein-like
LjNEST5A12	TC565	2.5	1.2E-02	Suppressor protein (ATP-binding)
LjNEST2A2	TC978	37.7	6.4E-05	Transcription factor
LjNEST53C2	S	3.1	4.5E-03	Transcription factor
LjNEST15H12	TC3127	2.8	8.2E-04	Transcription factor
LjNEST41A4	S	2.6	1.2E-02	Transcription factor
LjNEST6B9	S	2.5	1.4E-03	Transcription factor
LjNEST12A6	TC842	8.4	3.3E-03	Transcription factor IIB
LjNEST39C1	TC2149	4.0	3.1E-04	Transcription factor, APFI-like
LjNEST15F8	TC1357	8.7	4.8E-05	Trihelix DNA-binding motif
LjNEST24D8	TC1447	4.0	2.7E-03	WRKY family transcription factor
LjNEST1H12	S*	56.0	3.1E-05	Zinc-finger protein
LjNEST51B7	TC3683	4.6	2.3E-03	Zinc-finger protein
LjNEST11A11	TC3269	4.6	3.2E-03	Zinc-finger protein
LjNEST42C12	TC3685	2.9	2.0E-03	Zinc-finger protein
LjNEST53C8	TC1468	2.9	5.0E-03	Zinc-finger protein
LjNEST42B1	S	2.7	1.4E-03	Zinc-finger protein
LjNEST20D4	TC2030	2.2	3.1E-02	Zinc-finger protein

See Table S2 for an explanation of abbreviations.
*Previously found to be induced in nodules (Colebatch *et al.*, 2002).

Table 4 Nodule/root gene transcript ratios determined by real-time RT-PCR

EST/TC	Annotation	Array Nod. Sig.	Array Nod./Root	RT-PCR Nod./Root	SE ($n = 3$)
LjNEST44C3/TC2403	DNA-binding protein	0.03	3.00	2.01	0.04
LjNEST50D8/TC881	DNA-binding protein	0.04	3.20	1.38	0.23
LjNEST42C12/TC3685	Zinc-finger protein	0.06	2.90	0.54	0.05
LjNEST6F1/TC1391	bZIP transcription factor	0.06	3.40	1.26	0.06
LjNEST10E7/TC3007	Homeobox RRM-containing protein	0.07	2.30	0.83	0.06
LjNEST14B3/TC1629	KH domain/zinc finger protein	0.08	2.40	0.75	0.01
LjNEST42B1	Zinc-finger protein	0.08	2.70	0.96	0.19
LjNEST6A3/TC902	CCAAT-box-binding transcription factor	0.09	3.60	>8.33	0.81
LjNEST53C8/TC1468	Zinc-finger protein	0.09	2.90	0.51	0.01
LjNEST12E3/TC2417	DNA-binding protein	0.10	4.90	1.38	0.08
LjNEST24D8/TC1447	WRKY family transcription factor	0.10	4.00	0.50	0.05
LjNEST6B9	Transcription factor	0.11	2.50	2.05	0.04
LjNEST54F12/TC2072	DNA-binding protein	0.11	2.40	0.84	0.01
LjNEST16G12/TC1091	MYB family transcription factor	0.12	2.30	0.95	0.05
LjNEST46H8/TC3654	MADS-box protein	0.16	4.20	>168.85	3.48
LjNEST49H11/TC2380	Negative regulator of URS2	0.18	5.00	0.91	0.04
LjNEST53C2	Transcription factor	0.18	3.10	1.65	0.04
LjNEST55C5	PHD-type zinc finger protein	0.18	3.40	>267.44	16.58
LjNEST40D9/TC367	MYB family transcription factor	0.19	11.60	73.82	1.73
LjNEST11A11/TC3269	Zinc-finger protein	0.19	4.60	1.28	0.04
LjNEST25D11/TC507	Carbonic anhydrase	0.34	11.90	15.40	0.04
LjNEST3E1/TC2283	Carbonic anhydrase	0.73	84.20	41.78	0.28
LjNEST15A11/TC2173	Isoliquiritigenin 2′-O-methyltransferase	0.74	17.30	260.46	4.75
LjNEST12H12/TC2109	Nodulin, Nlj21	0.87	24.30	21.29	0.23
LjNEST12C4	Sulphate transporter	1.13	33.70	313.94	8.33
LjNEST16B9	GA 2-oxidase	1.37	39.50	48.66	0.37
LjNEST22F11/TC5829	Leghaemoglobin	21.40	214.60	623.99	10.51

The first 20 genes encode putative transcription factors, each with relative transcript level in nodules <0.2, as determined by cDNA array analysis (Array Nod. Sig.). The remaining seven genes were expressed at higher levels (Array Nod. Sig. >0.2). Nodule/Root transcript ratios obtained from cDNA array and RT-PCR analysis are shown, together with the standard error (SE) associated with the RT-PCR analysis. Transcripts of several genes were detected only in nodules by RT-PCR; nodule/root ratios for these genes are given as the theoretical minimum ratio (>), assuming the root value was at the detection limit (Ct = 40).

ficantly lower in nodules than roots, including those encoding: two homologues of pathogen-response protein, PR-10; two hydroxyproline-rich proteins; two lipoxygenases; a peroxidase; and a putative chitinase.

Metabolome analysis

Transcriptome analysis indicated that major shifts in plant metabolism occur during nodule differentiation. To assess the degree to which changes in plant gene expression affect overall metabolism, we performed non-biased metabolite profiling of *Lotus* organs, using gas chromatography coupled to mass spectrometry (GC-MS). First, we created a library of mass spectral metabolite tags (MSTs), to facilitate qualitative and quantitative comparisons between metabolite profiles of different organs. MSTs are analogous to ESTs in that each MST represents a unique metabolite, and the normalized signal associated with each MST provides a measure of the relative abundance of the matching metabolite in a biological sample.

GC-MS metabolite profiles allow non-biased collection of MSTs that are either ubiquitous throughout the plant or present only in specific organs (Wagner *et al.*, 2003). We created a non-supervised library containing 6527 MSTs obtained from two independent samples from each of the following six organs: primary roots, lateral roots, nodules, developing leaves, mature leaves, and flowers (File S1). This library, which contains mass-spectral information as well as gas chromatography retention time indices (RI) for identification purposes, was constructed using an automated and non-biased procedure and contains on average 544 MSTs per sample. Replicate analyses of each sample type allowed verification of MST occurrence in each organ and detection of errors of automated mass-spectral de-convolution. In a manner similar to a BLAST analysis of EST libraries, the non-supervised MST library was screened for known metabolites, which had been characterized previously by mass-spectral fragmentation and chromatographic retention in standard addition experiments (Wagner *et al.*, 2003). We identified 85 non-redundant MSTs (File S2), representing 71 metabolites (Table 5) by manual queries, performed using publicly available mass-spectral search and comparison software (see Experimental procedures). This software generated mass-spectral hit lists, which were analysed for the

Table 5 Metabolites identified by GC-MS in *Lotus japonicus* organs

	Mass to charge ratio (*m/z*)	Retention time index (RI), median	Retention time index (RI), standard deviation	Response ratio (nodule/root)	Response ratio (lateral root/primary root)	Component_1	Component_2	Component_3
Standard deviation						2.31	1.28	1.27
Proportion of variance (%)						33.67	10.33	10.12
Cumulative proportion (%)						33.67	44.00	54.12
A (amino acids)								
2-Aminoadipic acid	260	1728	1.3	**1.3**	**0.7**	–	–	–
4-Aminobutyric acid	304	1531	1.4	1.1	**0.2**	–	0.13	−0.13
β-Alanine	174	1432	0.8	**0.4**	**0.5**	–	0.14	−0.11
Glycine	248	1313	1.4	**3.6**	0.9	−0.10	0.15	0.19
L-Alanine	116	1095	3.1	**2.5**	1.0	–	–	0.15
L-Asparagine	116	1686	3.3	**5.9**	0.8	**−0.23**	−0.11	–
L-Aspartic acid	232	1526	1.8	1.2	**0.1**	−0.15	–	**−0.24**
L-Glutamic acid	246	1633	2.9	**9.5**	**0.1**	**−0.33**	−0.11	−0.14
L-Glutamine	156	1786	3.3	**14.3**	0.4	**−0.42**	**−0.28**	–
L-Isoleucine	158	1302	1.9	**2.3**	**0.5**	−0.13	0.22	–
L-Leucine	158	1278	4.0	**5.6**	**0.5**	−0.22	–	–
L-Lysine	156	1922	2.9	**1.7**	2.1	–	–	–
L-Ornithine	142	1822	2.0	2.1	**0.1**	−0.22	–	**−0.24**
L-Phenylalanine	192	1637	2.8	**2.1**	**0.2**	−0.18	–	−0.13
L-Proline	142	1304	1.4	**4.8**	**0.2**	**−0.29**	–	−0.22
L-Serine	204	1371	1.2	**2.0**	**0.2**	−0.16	0.21	–
L-Threonine	291	1395	1.4	0.8	**0.4**	–	0.18	–
L-Tryptophan	202	2217	3.9	0.2	1.4	0.15	–	−0.19
L-Tyrosine	280	1942	2.9	**1.1**	1.1	–	–	–
L-Valine	144	1221	1.7	**1.6**	1.0	–	–	–
Pyroglutamic acid	258	1528	1.9	**2.5**	**0.4**	−0.12	–	–
B (organic acids)								
2,3,4-Trihydroxybutyric acid (erythronic acid)	292	1550	1.3	**1.4**	1.2	–	–	–
2,3,4-Trihydroxybutyric acid (threonic acid)	292	1570	1.2	**18.4**	1.2	−0.20	–	**0.24**
Citramalic acid	349	1474	0.5	**0.2**	**0.6**	–	–	−0.18
Citric acid	273	1829	0.8	**0.3**	1.1	–	–	–
Dehydroascorbic acid dimer	316	1852	0.5	**2.1**	0.9	–	**0.25**	–
Fumaric acid	245	1363	0.9	**2.3**	**0.7**	–	–	–
Galactonic acid	292	1999	1.3	**2.6**	**0.3**	−0.13	0.13	–
Glucaric acid	292	2014	0.9	1.2	**0.5**	–	0.15	–
Gluconic acid	292	2003	2.3	**3.3**	**1.7**	–	–	**0.24**
Glutaric acid	158	1416	1.7	**3.0**	1.0	–	–	–
Glyceric acid	292	1341	2.0	0.5	1.0	–	0.19	–
Gulonic acid	333	1965	1.4	1.1	**0.7**	–	–	–
Hexadecanoic acid	313	2052	1.1	**1.6**	0.7	–	–	–
Lactic acid	219	1049	1.3	0.8	1.2	–	–	–
Maleic acid	245	1313	2.3	1.0	1.3	–	–	–
Malic acid	233	1493	1.9	**1.5**	0.9	–	–	–
Quinic acid	345	1862	1.1	**1.5**	0.9	–	–	–
Succinic acid	247	1327	1.5	1.2	**0.6**	–	–	–
Threonic acid-1,4-lactone	247	1385	1.7	**2.8**	**0.5**	−0.14	–	–
C (aromatic acids)								
Benzoic acid	179	1253	1.9	**1.4**	1.0	–	–	–
p-Aminobenzoic acid	281	1837	3.0	1.3	0.7	–	–	–
trans-p-Coumaric acid	308	1946	4.9	**0.5**	1.5	–	–	–
D (N-containing compounds)								
Allantoin	518	1905	4.5	0.9	0.7	–	–	–
Putrescine	174	1741	0.4	**3.0**	1.4	–	–	0.21

Table 5 continued

	Mass to charge ratio (*m/z*)	Retention time index (RI), median	Retention time index (RI), standard deviation	Response ratio (nodule/root)	Response ratio (lateral root/primary root)	Component_1	Component_2	Component_3
Urea	189	1270	2.5	0.9	0.9	–	–	–
Uric acid	441	2111	0.9	0.2	1.6	0.17	–	−0.19
E (sugars)								
Arabinose	160	1676	2.1	**0.5**	0.6	–	0.16	−0.19
Fructose	307	1875	0.7	**0.1**	**3.7**	0.22	0.19	−0.13
Fucose	160	1747	0.7	1.0	**0.6**	–	0.17	–
Galactose	160	1892	0.5	**0.5**	0.8	–	0.18	–
Glucose	160	1898	0.5	**0.4**	**4.1**	–	0.16	–
Maltose	160	2747	1.8	1.5	**2.1**	–	0.14	0.18
Raffinose	451	3401	3.0	**1.5**	**0.4**	–	–	–
Ribose	160	1691	1.8	**2.3**	1.4	–	–	0.17
Sucrose	437	2653	1.0	0.9	**0.6**	–	–	–
Trehalose	191	2751	2.2	**0.6**	0.6	–	–	–
Xylose	160	1670	1.3	**0.3**	2.2	–	**0.25**	−0.16
F (polyols)								
4-O-Methyl-myo-inositol, Ononitol	318	1955	0.6	**4.0**	**0.6**	−0.10	–	–
Erythritol	205	1511	3.1	1.0	1.3	–	–	–
Galactinol	204	2995	2.3	0.6	**0.2**	–	–	−0.19
Galactitol	307	1941	2.6	1.3	**0.8**	–	–	–
Glycerol	205	1278	2.8	1.0	0.9	–	–	–
Mannitol	319	1929	1.2	**3.0**	**2.1**	–	–	**0.23**
myo-Inositol	305	2091	0.4	1.1	**0.5**	–	0.14	–
Sorbitol	319	1937	1.2	**1.9**	1.5	–	0.13	0.18
Threitol	217	1503	3.9	0.8	1.0	–	–	–
G (phosphates)								
Fructose-6-phosphate	315	2324	2.4	**1.6**	2.3	–	–	–
Glucose-6-phosphate	387	2337	1.7	**3.0**	**0.8**	−0.12	–	–
Mannose-6-phosphate	160	2324	2.2	**2.6**	0.7	−0.13	–	–
Phosphoric acid	314	1282	1.0	1.3	**0.4**	–	–	–

The table lists the mass fragments and corresponding windows of retention time indices, which were used to retrieve metabolite response ratios from GC-MS profiles for principal component analysis of root and nodule samples. The first three components are characterized by standard deviation, proportion of variance, and metabolite loadings >|0.125|, respectively. Metabolites that contributed strongly to each of the components are highlighted in bold. The response ratio of average nodule response compared with average root response is listed, as is that for lateral to primary roots (*t*-test significance of $P < 0.05$ is indicated by bold format of the response ratio).

presence of MSTs of *L. japonicus* origin that were identical to MSTs generated from pure standard compounds. Criteria for manual validation of MST identification were mass-spectral match values >700 on a scale of 0–1000, where a match of 1000 indicates perfect identity, and occurrence within an RI window of ±2.5. The metabolites thus identified were mostly primary metabolites that belonged to the compound classes: amino acids, organic acids, sugars, sugar phosphates, polyols, and other nitrogenous compounds.

Quantitative data for the 71 metabolites listed in Table 5 were used for principal component analysis (PCA) to identify major differences in metabolite composition of roots and nodules. The relative amount of each compound in a sample, also called the metabolite response ratio, was calculated from GC-MS data as described in Experimental procedures. The MS fragment of each MST and the GC-RI window used to quantify each metabolite are listed in Table 5. PCA of metabolite response ratios of all 71 metabolites in 20 samples from primary roots, 25 samples from lateral roots, and 20 samples from nodules allowed non-biased partitioning into three distinct sample groups, which reflected clearly the respective origin of each sample (Figure 3). The first three components obtained from PCA accounted for 54% of the total variance observed within the whole data set. PCA provided insight into which compounds contributed most to the variance between organs. Metabolite loading data from PCA indicated that the following metabolites contributed most to inter-organ partitioning: asparagine, aspartic acid, glutamine, glutamic acid, and proline. Several sugars, polyols, and polyhydroxyacids,

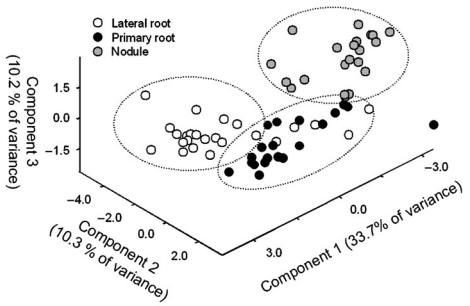


Figure 3. Principal component analysis of *Lotus* organ metabolite profiles. Principal component analysis of metabolite profiles from organs of nodulated *Lotus japonicus* plants.

such as fructose, ononitol, and galactonic acid, exhibited additional discriminatory power (Table 5). As expected, the nodule/root response ratios for these metabolites were further from unity than for other compounds (Table 5). Interestingly, concentrations of many amino acids, as well as the transport sugars sucrose, raffinose, and galactinol were significantly lower in lateral roots than in primary roots, which may reflect the central role of the primary root in long-distance transport of compounds and/or a greater metabolic sink for these compounds in lateral roots. Higher concentrations of fructose and glucose in lateral roots than in primary roots (Table 5) are not inconsistent with more rapid sucrose breakdown in the former.

Discussion

There has been an explosion in the numbers of EST sequences from legume species in the past few years. As of 2 April 2004, GenBank contained 346 582 ESTs from *Glycine max* (soybean), 187 763 from *Medicago truncatula* (barrel medic), and 110 563 from *L. japonicus*. Obviously, there are fewer genes in these species than the number of ESTs. One of the perceived benefits of redundancy in EST data sets is the insight that it can provide with respect to gene expression levels in different organs under different conditions. Thus, by comparing the frequency of occurrence of ESTs for a specific gene in different organs one can obtain a 'virtual Northern' for that gene. This approach was put to good use in identifying 340 genes in *Medicago* that appeared to be expressed in a nodule-specific manner (Fedorova *et al.*, 2002). Although not all of these could be confirmed experimentally as nodule-specific, by macroarray or RNA-blot analysis, many were confirmed. A similar *in silico* statistical approach was used to identify other nodule-induced genes in *Medicago* (Journet *et al.*, 2002). Together, these studies have uncovered many genes that were previously not known to be involved in nodule development or function, including many potentially involved in signal transduction, transcription, protein synthesis, and other cellular processes. However, relatively few genes involved in nodule metabolism were uncovered via the *in silico* approach, presumably because the majority of such genes are expressed throughout the plant, albeit at different levels in different organs under different conditions. Thus, the depth of EST sequencing is far from sufficient to yield statistically significant differences in gene expression for the majority of metabolic genes. Although alternatives to EST sequencing, such as Massively Parallel Signature Sequencing can generate millions of sequence tags rather than thousands for each organ and experimental condition, we chose to use cDNA arrays to compare the transcriptome of roots and nodules of *L. japonicus*. In this way, we were able to identify approximately 860 genes that were induced during nodule development, of which approximately one-third are involved in metabolism or transport. Fifty-seven of these genes were found to be nodule-induced previously (Colebatch *et al.*, 2002; see e.g. asterisks in Tables 1–3). Thus, 70% of the genes identified in our earlier work were confirmed as nodule-induced in the present study. Of the remaining genes, transcripts for some were found to be greater than twofold more abundant in nodules than in roots, but not at a statistically significant level ($P > 0.05$; LjNEST1E11, 2F7, 2H5, 11A7, 22B1; see Table S1), while others were significantly more highly expressed in nodules than in roots but did not reach the twofold induction threshold for reporting here ($P < 0.05$, ratio <2; e.g. LjNEST10a2; Table S1). Transcripts of other genes were not detected on the new arrays because of PCR failures during cDNA amplification prior to spotting (e.g. LjNEST1A8). PCR failures typically affect a few percentage of the clones spotted onto arrays.

The 860 nodule-induced genes identified here represent a 10-fold increase over the number we reported previously (Colebatch *et al.*, 2002). Part of this increase can be attributed to the greater number of genes represented on the new cDNA array. The rest may be explained by the modified data analysis and greater number of technical replicates used in the present study. In our earlier work, we excluded from analysis all clones/spots for which hybridization signals were less than twice the local background level (Colebatch *et al.*, 2002). Thus, data for many genes expressed at low levels were ignored. To access this data in the present study, we eliminated the twofold cutoff, and increased the number of technical replicates. Real-time RT-PCR confirmed the nodule-induced status of slightly more than half of the 20 lowest-expressed TF genes identified by cDNA array analysis (Table 4). These included four genes that were highly induced in nodules, which were not identified from the current data using our earlier data mining methods (Colebatch *et al.*, 2002). Obviously, however, extraction of such information came at the cost of a higher rate of false

positives. We include cDNA array data for genes expressed at very low levels in this report to facilitate future functional studies on important genes families, such as the TF genes, with the caveat that they should first be confirmed by real-time RT-PCR.

Global changes in gene expression during nodulation gear plant metabolism towards malate supply to bacteroids and asparagine synthesis and removal from nodules

We chose as the starting point for this work cDNA clones derived from a nodule EST project, which we expected would be biased for genes that are induced in nodules. Furthermore, because the main aim of this work was to identify changes in metabolism that follow nodule development, and because many metabolic genes are relatively highly expressed, we expected that a moderate number of EST clones, representing a few thousand genes, would provide a good overview of global changes in metabolism. Both of these assumptions were validated in the course of this work. Thus, although the majority of genes represented on the cDNA array were expressed at the same level in roots and nodules, approximately one-third of the genes were expressed at a significantly higher level in nodules. Few of the genes represented on the array were expressed at a lower level in nodules than roots. One-third of the nodule-induced genes, or approximately 10% of all the genes represented on the array are involved in metabolism and transport (Figure 2; Tables 1 and 2). Many of these genes encode enzymes that are part of well-characterized metabolic pathways, some of which have been studied in detail in legume nodules in the past.

More than 20 genes involved in sugar breakdown were more highly expressed in nodules than in roots (Table 1). These included genes for the majority of the enzymatic steps between sucrose and phosphoenolpyruvate (Figure 4). Sucrose, delivered from the shoot, is the primary source of carbon for nodule metabolism: a mutation in the *rug4* gene of pea, which encodes a nodule-induced form of sucrose synthase, severely impairs nodule functioning and SNF (Craig *et al.*, 1999; Gordon *et al.*, 1999). Although nodules of several legume species have enhanced activity of many of these enzymes, compared with roots (Copeland *et al.*, 1989, 1995; Day and Copeland, 1991; Gordon and James, 1997), until now the basis for these changes was largely unknown. Our data indicate that for many of the enzymes, increased activity is programmed at the level of transcription and/or RNA stability. This was previously known only for sucrose synthase in different legumes (Colebatch *et al.*, 2002; Perlick and Puhler, 1993; Thummler and Verma, 1987), for enolase in *Alnus glutinosa* (van Ghelue *et al.*, 1996), and for enolase and cytosolic phosphoglycerate kinase in *Lotus* (Colebatch *et al.*, 2002). The significant decrease in *Lotus* nodule/root ratios for the hexoses, fructose and glucose, and the concomitant increase in nodule/root ratios of hexose phosphates, and of alanine and serine (Table 5), is indicative of greater glycolytic flux in nodules compared with roots (Fernie *et al.*, 2002; Roessner *et al.*, 2001). A significant increase in the concentration of sugar alcohols in nodules compared with roots (Table 5) is also consistent with increased polyol biosynthesis in nodules, which in the case of sorbitol appears to be programmed at the level of transcription (Table 1, sorbitol dehydrogenase).

Two novel nodule-induced genes encoding starch phosphorylases were identified in this study (Table 1). One of these (TC1455) is homologous to SEX1/R1, which is required for starch degradation in plants (Ritte *et al.*, 2002; Yu *et al.*, 2001). High activity of these proteins in nodules may ensure rapid turnover of starch, and account for the lack of starch accumulation in this organ. Starch accumulation in amyloplasts is observed only in defective nodules that are not a strong sink for carbon, such as those induced by *nif*⁻ rhizobia or those produced by plant *sym* mutants (Tansengco *et al.*, 2003; Vance and Johnson, 1983).

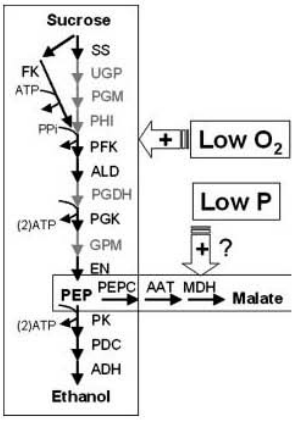


Figure 4. Model showing possible induction of nodule cell glycolysis and CO_2 fixation by low oxygen and low phosphorous, respectively.
Enzymes, transcripts of which were significantly more abundant in nodules than in roots, are indicated with black text/arrows, and include: SS, sucrose synthase; FK, fructokinase; PFK, phosphofructokinase; ALD, aldolase; PGK, phosphoglycerate kinase; EN, enolase; PK pyruvate kinase; PDC, pyruvate decarboxylase; ADH, alcohol dehydrogenase; PEPC, phosphoenolpyruvate carboxylase; AAT, aspartate aminotransferase; MDH, malate dehydrogenase. Transcript levels of the following enzymes (grey text/arrows) were unaltered in nodules compared with roots: UGP, UDP glucose pyrophosphorylase; PGM, phosphoglucomutase; PHI, phosphohexose isomerase; PGDH, phosphoglyceraldehyde dehydrogenase; and GPM, phosphoglycerate mutase. Low oxygen concentration induces glycolysis, which yields a surplus of ATP via substrate-level phosphorylation. Note that two nodule-induced pyrophosphate-dependent phosphofructokinases reduce the ATP requirement of glycolysis. The model also proposes that low P availability may induce carbon fixation, leading to increased malate synthesis in nodules compared with roots (see also Table 5).

Malate, rather than sugars, is likely to be the primary source of carbon for bacteroid metabolism and SNF (Day and Copeland, 1991; Streeter, 1995), and transporters that deliver dicarboxylates to the bacteroids have been characterized biochemically (Udvardi *et al.*, 1988). Direct evidence for the central role of dicarboxylates in bacteroid SNF came from genetic studies of rhizobial mutants impaired in transport or metabolism of these compounds (Arwas *et al.*, 1985; Gardiol *et al.*, 1982; Ronson *et al.*, 1981). Malate is produced from phosphoenolpyruvate (PEP) via PEP carboxylase (PEPC) and malate dehydrogenase, both of which are induced during nodule development in other species (Hata *et al.*, 1998; Imsande *et al.*, 2001; Miller *et al.*, 1998; Pathirana *et al.*, 1992; Suganuma *et al.*, 1997). We identified two PEPCs and two CAs, which feed CO_2 to PEPC via bicarbonate, as nodule-induced in *Lotus* (Table 1). Similar results were published recently for PEPC from *Lotus* (Colebatch *et al.*, 2002; Nakagawa *et al.*, 2003). Induction of CA genes in nodules has also been described in other legume species (Kavroulakis *et al.*, 2000; de la Pena *et al.*, 1997). Two malate dehydrogenase genes were found to be induced threefold in nodules compared with roots of *Lotus* (Table 1). Significant increases in the concentration of malate in nodules compared with roots were measured in *Lotus* (Table 5), which presumably reflects the increased transcript levels for enzymes involved in the synthesis of this important compound (Figure 4). A number of other genes involved in carbon fixation and dicarboxylate or glyoxylate metabolism were induced in *Lotus* nodules (Table 1). Interestingly, one of these encodes ribulose-bisphosphate carboxylase (Rubisco) small unit, which is normally associated with photosynthesis. A homologue of this gene was one of the few metabolic genes to be identified by *in silico* analysis of the *Medicago* nodule transcriptome (Fedorova *et al.*, 2002). A Rubisco activase gene was also found to be upregulated in *Lotus* nodules compared with roots (Table 1). The role, if any, of Rubisco in nodules is unknown.

Numerous genes involved in ammonium assimilation and asparagine synthesis, including two encoding GS, seven for AAT, and two for asparagines synthases (AS) were induced in nodules (Table 1). Asparagine is the major transport form of N exported from *Lotus* nodules (Vance *et al.*, 1987), and upregulation of GS, AAT, and AS presumably ensures rapid synthesis of this amino acid. Nodule-induced homologues of these genes have been identified in other legumes (Cullimore *et al.*, 1984; Gregerson *et al.*, 1994; Lara *et al.*, 1983; Perlick and Puhler, 1993; Reynolds *et al.*, 1992; Shi *et al.*, 1997; Tsai and Coruzzi, 1990). In the case of AAT, however, it has never before been reported that so many gene family members are upregulated during nodulation. An immediate concern with our data was that cross-hybridization between family members could account for the apparent upregulation of so many different genes. However, the data from transcriptome analysis (Table 1) are largely consistent with the frequency of occurrence of ESTs of these genes in different cDNA libraries. For five of the seven putative AAT genes, ESTs were present exclusively in nodule libraries; while for TC500, two-thirds (4/6) of all ESTs were present in nodule libraries although nodule ESTs represent a minority of all *Lotus* ESTs currently in the public domain. It will be interesting to determine where these genes are expressed within nodules, and where the encoded proteins are located within cells in order to understand better their individual roles in nitrogen assimilation in this organ. Metabolite profiling by GC-MS supported the conclusion from transcriptome analysis that biosynthesis of the amides glutamine and asparagine is enhanced in nodules (Table 5). In fact, these two metabolites together with glutamic acid and proline, which were also more concentrated in nodules, proved to be most powerful in discriminating between the different underground organs in PCA analysis (Table 5).

Over 30 other genes involved in amino acid metabolism were found to be nodule-induced (Table 1). Interesting amongst these were genes involved in proline and polyamine synthesis. These included a gene encoding ornithine cyclodeaminase, which converts ornithine to proline, and two genes for proline oxidases, which carry out a reversible reaction between 1-pyrroline-5-carboxylate and proline. In soybean, a gene encoding pyrroline-5-carboxylate reductase (P5CR), which also produces proline, was found to be both nodule-induced and osmoregulated (Delauney and Verma, 1990). Proline is a compatible solute that accumulates in plant tissues, including nodules, in response to osmotic stress (Liu and Zhu, 1997; Swaraj and Bishnoi, 1999). Other compatible solutes include sugar alcohols and polyamines. GC-MS revealed accumulation of a number of these compounds in *Lotus* nodules, including proline, ononitol, mannitol, sorbitol, and putrescine (Table 5). Enhanced expression of genes for ornithine decarboxylase and arginine decarboxylase, which are involved in polyamine biosynthesis, presumably contributed to increased putrescine biosynthesis in nodules (Table 1). Although polyamine composition varies considerably between species, legume nodules typically contain higher concentrations of these compounds than is found in other organs (Fujihara *et al.*, 1994). Thus, legume nodules may be subject to osmotic stress even under normal conditions.

Nodule-induced transporters

Legume nodules are highly specialized organs and it is assumed that many transporter genes are induced during nodule differentiation in order to handle the high fluxes of sugars into, and amino acids and other nitrogen compounds out of this organ. It is also likely that many of the proteins of the SM, which mediate transport of nutrients between the

plant and rhizobia are encoded by genes that are induced or expressed specifically in nodules. Nonetheless, few nodule-induced transporters have been identified in the past, and even fewer have been characterized in detail. The latter include the soybean SM proteins nodulin 26, an aquaporin (Fortin *et al.*, 1987; Rivers *et al.*, 1997; Weaver *et al.*, 1994), GmZIP1 a zinc transporter (Moreau *et al.*, 2002), and GmSAT1 (Kaiser *et al.*, 1998). The apparent *Lotus* orthologue of nodulin 26, LIMP 2 is also a nodulin, and a multifunctional aquaglyceroporin (Guenther and Roberts, 2000). Other nodule-induced transporters include *Lotus* nodulin LjN70, which is homologous to oxalate/formate exchange proteins (Szczyglowski *et al.*, 1998) and soybean nodulin Gm70, which is a putative sulphate transporter (Kouchi and Hata, 1993). Previous transcriptome analysis in *Lotus* uncovered a further four nodule-induced transporter genes, encoding two putative sulphate transporters, a potassium transporter, and a MATE efflux protein (Colebatch *et al.*, 2002). *In silico* analysis of EST databases revealed nine nodule-induced transporter genes in *M. truncatula* (Fedorova *et al.*, 2002).

The transcriptome analysis presented here uncovered 46 nodule-induced transporter genes, which is more than the total number described previously for all legumes combined (Table 2). Three genes encoding putative sugar transporters were identified amongst the nodule-induced genes. These may be important in cellular uptake and distribution of sugars throughout the nodule. Previous work indicated that hydrolysis of sucrose, which enters nodules via the phloem, occurs predominantly in uninfected, inner cortical cells of mature nodules (Day and Copeland, 1991). Recent work on *LjSUT4*, which encodes a putative sucrose transporter and was identified as a result of our nodule EST project, indicates that it is expressed in the vascular bundles and inner cortex of mature *Lotus* nodules (Flemetakis *et al.*, 2003). Thus, LjSUT4 may play an important role in delivering sucrose to the cells that are the major metabolic sink for this sugar. *LjSUT4* was not amongst the genes we found to be nodule-induced. Instead, three homologues of genes that have been variously annotated as hexose or mannitol/sorbitol transporters were identified (Table 2). A homologue of these genes was recently reported to be nodule-induced in *Medicago* (Fedorova *et al.*, 2002). It will be fascinating to determine the substrate specificity of these transporters, their location within nodule cells, and their disposition relative to the many nodule-induced glycolytic enzymes identified here (Table 1).

A gene encoding a putative cationic amino acid transporter of the APC superfamily (amino acid-polyamine-organocation) was also found to be nodule-induced (Table 2). As the name implies, this transporter may be involved in amino acid and/or polyamine transport. A homologue of this gene was also identified as nodule-induced in *Medicago*, by *in silico* analysis (Fedorova *et al.*, 2002). It will be interesting to determine whether the substrates of these proteins include glutamine, a major export form of N from nodules, or polyamines, which are also produced in relatively high amounts in nodules (see above). Other genes that may be involved in N-transport in nodules include five encoding proteins of the PTR/POT family, which transport peptides, amino acids, and/or nitrate, and one OPT family member, which may transport oligopeptides (Stacey *et al.*, 2002). It will be interesting to determine whether these nodule-induced proteins play a role in mass-flow of N in nodules, and/or or in signalling processes that may involve peptides.

Signalling and transcription: suspects galore

A plethora of putative signalling and regulatory genes was induced in mature nodules of *Lotus* (Tables 3 and 4). These included numerous kinases and phosphatases. The phosphorylation status of many enzymes and transporters is likely to be important for regulation of nodule metabolism and nutrient exchange between the plant and bacteroids. For instance, the nodule-enhanced PEPC is subject to post-translational regulation by PEPC kinase *in vitro* and *in vivo* (Schuller and Werner, 1993; Zhang and Chollet, 1997a,b), and both the kinase and its target enzyme are upregulated in soybean nodules (Wadham *et al.*, 1996; Zhang and Chollet, 1997a). Phosphorylation of PEPC renders it less sensitive to feedback inhibition by malate (Schuller and Werner, 1993), which may ensure constant synthesis of malate for SNF, even in the face of high cytoplasmic concentrations of this compound. GC-MS measurements revealed significantly higher concentrations of malate in nodules than in roots (Table 5). Until recently, it was not known whether increased activity of PEPC kinase in nodules was programmed at the level of transcription. Our data indicate that this is indeed the case in *Lotus* nodules (Table 3), and similar data have now been published for the orthologue in soybean (Xu *et al.*, 2003). Nodule-induced sucrose synthase is another target of phosphorylation in legumes (Komina *et al.*, 2002; Zhang and Chollet, 1997c), although the regulatory significance of this remains unclear.

At least one nodule transporter, the aquaporin nodulin 26 from soybean, is regulated by phosphorylation (Lee *et al.*, 1995), which is achieved by a calcium-dependent protein kinase (CDPK) in the SM (Weaver and Roberts, 1991; Weaver *et al.*, 1991). The *Lotus* orthologue of Nodulin 26, LIMP 2 is also phosphorylated by a CDPK (Guenther and Roberts, 2000). The nodule CDPK responsible for this phosphorylation remains to be identified. Interestingly, transcriptome analysis identified a CDPK gene that was highly upregulated in *Lotus* nodules (Table 3).

Receptor-like kinases have taken centre stage in symbiosis research recently, because of discoveries implicating them in local and long-distance regulation of nodule development (Endre *et al.*, 2002; Krusell *et al.*, 2002; Nishimura *et al.*,

2002; Searle *et al.*, 2003; Stracke *et al.*, 2002). Nothing is known about the roles of members of this family in mature nodules. We identified several genes encoding putative receptor kinases that exhibited elevated transcript levels in mature nodules compared with roots (Table 3), which may play crucial roles during SNF.

Homologues of other regulatory proteins that play important roles in plant–microbe interactions were also found to be nodule-induced (Table 3). Notable amongst these were two homologues of plant MLO proteins, which interfere with plant cell death and the onset of defence responses (Buschges *et al.*, 1997). It is not known how plants avoid triggering defence responses against beneficial microbes such as rhizobia (Mithofer, 2002). The *Lotus* MLO homologues are interesting suspects as suppressors of plant defence responses in nodules.

In silico screening of *Medicago* EST libraries led to the identification of a family of six nodule-specific genes encoding calmodulin-like proteins (Fedorova *et al.*, 2002). Several homologues of these genes were found to be upregulated in *Lotus* nodules (Table 3). Calcium is an important second messenger in plant cells, and calmodulins play essential roles in mediating calcium signalling (Reddy, 2001; Snedden and Fromm, 1998; Zielinski, 1998). Fedorova *et al.* (2002) highlighted the presence of a predicted N-terminal extension on calmodulin-like proteins of *Medicago*, which might direct these proteins out of the cytoplasm and possibly into the symbiosomes. The presence of high concentrations of calcium-binding proteins in symbiosomes could account, in part, for the accumulation of calcium that has been observed there by electron microscopy (Izmailov *et al.*, 1999). Although the role of such proteins remains to be determined, they may be crucial to calcium homeostasis and signalling in nodule cells.

The massive changes in gene transcription that occur during *Lotus* nodule development and differentiation documented here imply the involvement of many TFs to orchestrate the changes in a coordinated manner. TF genes account for at least 5% of plant genomes (Riechmann *et al.*, 2000), although the role of TFs in nodule development and differentiation is virtually unexplored territory. We identified many putative TF genes that are more highly expressed in nodules than roots (Tables 3 and 4). So far, only two putative TFs have been found to have crucial roles in SNF: NIN from *L. japonicus* (Schauser *et al.*, 1999); and Mszpt2-1 from alfalfa (Frugier *et al.*, 2000), which are required for nodule organogenesis and differentiation, respectively. Apart from these genes, several nodule-induced MADS-box genes have been identified in alfalfa, although their roles remain unknown (Heard and Dunn, 1995; Zucchero *et al.*, 2001). A DNA-binding protein, ENBP1 that binds to the promoter of ENOD12 from pea has also been identified (Christiansen *et al.*, 1996), but again the *in vivo* role of this protein remains to be determined.

Physiological Insights from molecular watchdogs

Gene expression and enzyme activity profiles have long been used as a 'barometer' of the physiological conditions that rhizobia encounter within nodules. For instance, activity of the oxygen-labile enzyme, nitrogenase in bacteroids implies that local oxygen concentrations in nodules are extremely low. In fact, free-oxygen concentrations are generally below 50 nM in the infected zone of nodules (Bergersen, 1997), triggering expression of the nitrogenase genes, and many others that are essential for SNF, which are under tight oxygen control (Batut and Boistard, 1994; Fischer, 1996). Likewise, measurements of the activity of enzymes and genes involved in rhizobial carbon and nitrogen metabolism, together with the use of key mutants, indicate that dicarboxylates, rather than sugars or amino acids, are the principal source of carbon for bacteroids (Arwas *et al.*, 1985; Gardiol *et al.*, 1982; Ronson *et al.*, 1981). Similar data also led to the surprising conclusion that bacteroids are not nitrogen-starved in nodules, and that nitrogenase is expressed for another reason, possibly as a means to ensure carbon supply from the plant (Udvardi and Kahn, 1993; Lodwig *et al.*, 2003).

Transcriptome analysis of thousands of genes provides not only the opportunity to identify genes and pathways that are involved in a specific biological process such as SNF, but also a window onto the physiological conditions that prevail within an organism. Specific external or internal physiological cues often lead to characteristic responses at the molecular level, and the recent use of DNA arrays has added significantly to our knowledge in this regard (Klok *et al.*, 2002).

Plant roots respond to hypoxic conditions, during flooding for instance, by inducing expression of genes involved in a number of processes, including glycolysis, fermentation, and ethylene biosynthesis (Drew, 1997; Geigenberger, 2003). Hypoxia induces expression of at least 20 anaerobic proteins (ANPs) in maize roots, most of which are enzymes of glycolysis and fermentation, or sugar–phosphate metabolism (Sachs *et al.*, 1996). Glycolysis and fermentation represent an alternative, albeit less efficient way to generate ATP for cellular processes when oxidative phosphorylation is restricted or stopped. Hypoxic conditions prevail in cells within the central, infected zone of nodules (Bergersen, 1997; Tjepkema and Yocum, 1974), and the increase in transcript levels in nodules compared with roots for glycolytic enzymes (Table 1) may well be a result of signals related to hypoxia. Hypoxia also results in increased activities of ACC synthase and ACC oxidase, and higher concentrations of ACC and ethylene in roots, which has been linked to aerenchyma formation (Drew, 1997). Development of aerenchyma is a long-term adaptation to hypoxia that increases gaseous diffusion within plant tissues, including nodules (Parsons and Day, 1990). Intriguingly, although no

ACC synthase genes were represented on our arrays, cDNA for several different ACC oxidases were present, two of which were induced significantly in nodules compared with roots (TC223 and LjNEST6h2).

Recent transcriptome analysis in *Arabidopsis* confirmed past work on low-oxygen responses in plants, and extended it especially in the realm of hypoxia signal transduction (Klok *et al.*, 2002). Several genes encoding putative signalling components were induced in hypoxic root cultures, including TFs of the MADS, WRKY, AP2, and MYB families, a putative RNA binding protein, and several types of kinases, including MAPK and receptor-like kinases. Interestingly, homologues of some of these genes were induced in *Lotus* nodules, for example, the MADS-box protein TC3654, the MYB TC367; RNA-binding protein TC2072; and the kinases TC3045, TC2786, and LjNEST56g5 (Tables 3 and 4). After 20 h of hypoxia, genes involved in protein ubiquitylation were induced in *Arabidopsis* roots (Klok *et al.*, 2002), and a similar situation existed in mature *Lotus* nodules (Table S2). This suggests that selective protein turnover may be an important aspect of regulation during sustained microaerobiosis.

Evolution has endowed plants with an array of adaptive responses to low phosphorous, which are manifest at different levels: morphological, physiological, and biochemical (Raghothama, 1999). At the biochemical level, production and excretion of organic acids, especially malate and citrate, into the rhizosphere under P-limiting conditions displaces inorganic phosphate (Pi) from soil particles, making it available for uptake by high-affinity Pi transporters in the plasma membrane of root cells. Increased activities of PEPC and malate dehydrogenase in roots in response to P-limitation have been recorded in many plants, including legumes (Uhde-Stone *et al.*, 2003). Increases in the activity of high-affinity Pi transporters in roots have also been measured under similar conditions (Mimura, 1999). P-deprivation also triggers increased production of intracellular and extracellular phosphatases for mobilization of organic P in the cell and rhizosphere, respectively (Raghothama, 1999). Remobilization of organic P within the cell is achieved via breakdown of nucleic acids, nucleotides, and phospholipids (Dormann and Benning, 2002; Raghothama, 1999). Substitutes for these important compounds include galactolipids and sulpholipids for phospholipids (Dormann and Benning, 2002), and possibly pyrophosphate in place of ATP (Raghothama, 1999). The majority of these biochemical changes are programmed at the transcriptional level (Raghothama, 1999; Uhde-Stone *et al.*, 2003). Therefore, it is salient to note that genes encoding homologues of all of these enzymes and transporters were induced in nodules (Tables 1 and 2). As mentioned above, two PEPC and two malate dehydrogenase genes were induced in nodules, as was a high-affinity Pi transporter homologue. Numerous phosphatase and lipase genes, as well as others involved in galactolipid and sulpholipid synthesis were also nodule-induced. At least one acid phosphatase is highly upregulated during soybean nodule development (Penheiter *et al.*, 1997), and evidence for enhanced levels of galactolipids in *Lotus* nodules has also been obtained (Peter Dörmann, personal communication). Taken together, these data indicate strongly that plant cells within legume nodules experience profound P-limitation. This is in apparent contradiction to measured phosphate levels in nodules, which indicate that free phosphate (phosphoric acid) concentrations in this organ are not less than in roots (Table 5). A resolution to this apparent paradox may come from precise intracellular measurements of P in nodules. One intriguing possibility is that bacteroids steal much of the nodule P away from the plant. Indeed, bacteroids express a high-affinity Pi transporter, which is essential for effective SNF (Bardin *et al.*, 1996). By removing Pi from the rest of the plant cell, bacteroids may trigger synthesis of malate, the principal source of carbon for SNF. A model showing how low cytoplasmic Pi together with low free O_2 in *Lotus* nodule cells may regulate plant glycolysis and CO_2 fixation to enhance malate supply for SNF is presented in Figure 4.

In summary, we have presented the first global overview of plant metabolic differentiation during legume nodulation that combines data from both transcriptome and metabolome analyses. The results presented here confirm and extend significantly data from many legume species scattered in a large number of publications over many years. By focusing the tools of functional genomics on a single model species in a non-biased way, a coherent picture of metabolism has emerged that provides new insights into SNF and possible players in its regulation.

Experimental procedures

Biological materials

Lotus japonicus GIFU (*B-129*) seeds were scarified in liquid nitrogen (3 × 10 sec), sterilized in a 2% bleach solution for 10 min, rinsed five times with sterile distilled water, then germinated and grown in coarse quartz sand in a controlled environment (16 h day, 60% relative humidity, and 21/17°C day/night temperate regime). Pots were watered with 1/4 B&D medium (Broughton and Dilworth, 1971). Plants were inoculated when 7 days old with *Mesorhizobium loti* strain R7A and provided with 1 mM KNO_3 for the first 2–3 weeks of growth. Plant organs were harvested directly into liquid nitrogen, and stored at –80°C.

Construction of nodule cDNA libraries

Two *L. japonicus* nodule cDNA libraries were used to generate clones for array analysis (Colebatch *et al.*, 2002). A total of 9600 clones from the two libraries were sequenced from the 5′ end at AGOWA (Berlin, Germany) and 8460 high-quality sequences were deposited into Genbank.

Construction and use of cDNA arrays

Second-generation arrays containing 5376 PCR-amplified cDNA clones were produced essentially as described by Colebatch *et al.* (2002). Each clone was spotted twice on every array. Reference hybridizations to determine the amount of cDNA applied to each spot on each filter were performed using ^{33}P-labelled vector-specific oligonucleotides (Colebatch *et al.*, 2002). Hybridization of ^{33}P-labelled first-strand cDNA, derived from nodules and uninfected roots of 7-week-old plants, was also performed as described previously (Colebatch *et al.*, 2002). At least five replicate hybridizations, representing two independent experiments (biological replicates) were performed for each organ. In the case of root transcriptome analysis, data from three hybridizations for one biological replicate and two hybridizations for the second replicate were analysed. For nodule transcriptome analysis, three hybridizations were performed for one and five for the second biological replicate.

cDNA array data analysis

Detection and quantification of signal intensities was as described previously (Colebatch *et al.*, 2002), with the following exception. Spot signal intensities resulting from hybridizations with cDNA probes that were less than twice the local background intensity (obtained from a neighbouring empty spot) were not automatically regarded as undetectable and set to zero, as set previously. Instead, the value for the local background was simply subtracted from the signal of each spot in the 4×4 subgrid. An average gene activity was accepted when positive non-zero values were obtained from at least three of the filters, with at least one from each of the two biological replicates. The Student's *t*-test was used to identify significant differences between the mean normalized gene activity in roots and nodules.

Real-time RT-PCR analysis of transcript levels

Quantitative determination of relative transcript levels using real-time RT-PCR was carried out according to Czechowski *et al.* (2004), except that total RNA (1 μg) was used instead of poly(A)+ RNA as template for cDNA synthesis. Gene-specific primers were designed using Primer Express Software (Applied Biosystems, Foster City, CA, USA). The constitutively expressed ubiquitin gene (TC3806) was used to normalize transcript data (Wandrey *et al.*, 2004). The sequences of oligonucleotide primers used in RT-PCR experiments are shown below (Table 6).

Plant metabolite extraction and derivatization

Nodules and roots were harvested from 12-week-old plants to obtain sufficient material for multiple technical replicates of each organ from pooled plants. Care was taken to pick pink nodules of a similar size range as those taken from the 7-week-old plants used in transcriptome analysis. Senescent nodules were excluded from both metabolome and transcriptome analyses. For each sample for metabolite analysis, between 25 and 50 mg (FW) frozen tissue was pulverized in a 2 ml Eppendorf tube containing a clean stainless steel metal ball (5 mm diameter) in a mixer-mill grinder (MM200; Restch, Haan, Germany) for 2 min at 30 cycles sec^{1}. Grinding components of the mill were cooled with liquid nitrogen to keep the

Table 6 Primers used for real-time RT-PCR

Target gene (EST/TC)	Forward primer	Reverse primer
LjNEST44C3/TC2403	5′-CACCGATACGGTGATGGAATCT-3′	5′-CGGCACCATTTCTTGCTGAT-3′
LjNEST50D8/TC881	5′-GTGCAATGCAAGGTTGGCA-3′	5′-TAACTGGCTCAGCAACCTCACC-3′
LjNEST42C12/TC3685	5′-TGTTTGGACACATTGCCCG-3′	5′-CCCCTATTTCGTCCAACGCTAT-3′
LjNEST6F1/TC1391	5′-AGCTCCACATGAATCCACACC-3′	5′-GGTGGCTAGCTTGGAACATCTT-3′
LjNEST10E7/TC3007	5′-TTGACCCTTCCTTCGGAATCA-3′	5′-TCAGTTGCCCACAGAACCTGA-3′
LjNEST14B3/TC1629	5′-GCACTTTTGGAGAACGATGCC-3′	5′-CCACTCCTCATTTCCGCAATT-3′
LjNEST42B1	5′-CCTTTGCCAATCCTGGTTCT-3′	5′-GCAACCTCTTCTTGCTCATGTT-3′
LjNEST6A3/TC902	5′-ACGATTCAAGAGTGCGTGTCG-3′	5′-ACCGCTCATTAGCTTCGGAAG-3′
LjNEST53C8/TC1468	5′-CGATGCAGGATCTCTGCTCAA-3′	5′-CGGAGGAAGATTGTTGCTCGT-3′
LjNEST12E3/TC2417	5′-CAGTTGGAGCTTTGCCACCTAG-3′	5′-TGTGAGAGCAGCCTTCCCTTT-3′
LjNEST24D8/TC1447	5′-TTCCCTCAAATGGACCAACAGA-3′	5′-CCATGCCTTTTAAGCCTTCGTT-3′
LjNEST6B9	5′-AGGTATTGAATGCCAACAGTGC-3′	5′-AACCCATTACTTGGCTCAGTGA-3′
LjNEST54F12/TC2072	5′-AGGTTGTGTCAGCGAAGGTCA-3′	5′-CAAATCCATATCCCTCCGACTG-3′
LjNEST16G12/TC1091	5′-AGCCGAGGATCCGAGTAAGAAG-3′	5′-CAGCTCTCCCTGGACTTGGTAA-3′
LjNEST46H8/TC3654	5′-TTGTCTTCTCGCCAGGTGAAAA-3′	5′-TGATGACGGCCTCAACACTTG-3′
LjNEST49H11/TC2380	5′-GGAGCTGATGCTTCTATGCAGG-3′	5′-GCCCAAGGTTTTCTGAATGGA-3′
LjNEST53C2	5′-AAACCTTTCTCAGCCCAGCATT-3′	5′-TTGCCATCCATCACCATTCC-3′
LjNEST55C5	5′-CCAAGTGGTTTGTGGTACTGCA-3′	5′-CCCCTGGTCCACATCTTTGTT-3′
LjNEST40D9/TC367	5′-CCAACAACTGCTGATGAGTGGA-3′	5′-CATATGATGTCCCACTGCAGGA-3′
LjNEST11A11/TC3269	5′-GATGCTGAGGATCAGCCTTCTG-3′	5′-CCCATCCTCCTGTTCTGCAAT-3′
LjNEST25D11/TC507	5′-TTCTACAAACGCCACGCTTGT-3′	5′-TGTCCTTGAAAATGCACCCC-3′
LjNEST3E1/TC2283	5′-TCCCTAGTGAACCTGCACACAT-3′	5′-AAAGAGCCCTAACCCCTTGCT-3′
LjNEST15A11/TC2173	5′-CTGCCAGAAGATCCAGAAACTAC-3′	5′-CCAGAGCGCTTGCCCAAACACTC-3′
LjNEST12H12/TC2109	5′-TGCATCAGCAAACTCTGCC-3′	5′-CTCATGATCCTTCAATGAAGCC-3′
LjNEST12C4	5′-TCTTGGACTGGTCTTCCTTGCT-3′	5′-TGGTCTCTTGTTCCTCACGTGT-3′
LjNEST16B9	5′-GACATGCAAGCCCACAAACT-3′	5′-CTATGAGTGTTTTGGCCTCAGG-3′
LjNEST22F11/TC5829	5′-CTCCAAGCCCATGCTGAAAA-3′	5′-TGGCATCTGCAAGTGTCACTTC-3′
Ubiquitin TC3806	5′-TTCACCTTGTGCTCCGTCTTC-3′	5′-AACAACAGCACACACAGACAATCC-3′

sample as cold as possible. Metabolites were extracted according to (Roessner *et al.*, 2000) with the following modifications. Ground samples were mixed with 360 µl methanol (–20°C) plus 30 µl ribitol in methanol (0.2 mg ml^{-1}) and 30 µl deuterated d_4-alanine in water (1 mg ml^{-1}). Samples were shaken for 15 min at 70°C before addition of 200 µl chloroform and further shaking at 37°C for 5 min. After addition of 400 µl water, samples were vortexed, then centrifuged at 21 000 *g* for 5 min at RT. Two 80 µl aliquots of the aqueous phase were transferred to Eppendorf tubes and dried at RT by vacuum centrifugation (Centrivac; Hereaus, Hanau, Germany).

Metabolite derivatization was performed according to (Roessner *et al.*, 2000) with the following modifications. One of each pair of dried samples was re-suspended in 40 µl methoxyaminhydrochloride (MEOX, 20 mg ml^{-1} in pyridine) at 30°C for 90 min; 70 µl N-methyl-N-(trimethylsilyl)-trifluoroacetamide (MSTFA) plus 10 µl alkane mixture (see below).

GC-MS metabolite profiles

GC-MS spectra were obtained with a GC8000 gas chromatograph coupled to a Voyager quadrapol-type mass spectrometer, operated by MassLab software (ThermoQuest, Manchester, UK). Modifications to the initial GC-MS profiling method (Fiehn *et al.*, 2000a,b) included injection of a 1-µl sample in split-less mode, use of a 5°C min^{-1} temperature ramp with final temperature set to 320°C on a 30 m × 0.25 mm inner diameter Rtx-5Sil MS capillary column with an integrated guard column (Restek GmbH, Bad Homburg, Germany), and use of the C_{12}, C_{15}, C_{19}, C_{22}, C_{28}, C_{32}, C_{36} *n*-alkane mixtures for the determination of RI.

Compendium library of mass spectral metabolite tags

Mass spectral metabolite tags were obtained by automated de-convolution of GC-MS chromatograms using publicly available software (AMDIS, http://www.chemdata.nist.gov/mass-spc/amdis/; National Institute of Standards and Technology, Gaithersburg, MD, USA) (Stein, 1999). Mass spectra were collected above a threshold of 0.001% of total signal. Two independent GC-MS metabolite profiles of each sample type were processed. RIs of all MSTs were determined by AMDIS software. RI-annotated MSTs were uploaded into a non-supervised custom NIST02 mass spectral library (NIST02 mass spectral search program, http://www.chemdata.nist.gov/mass-spc/Srch_v1.7/index.html; National Institute of Standards and Technology, Gaithersburg) (Ausloos *et al.*, 1999). Mass spectra of low relative amount were rejected. The resulting MST library of *L. japonicus* root and shoot organs is available as Supplementary material, NIST02 custom library file Q_LJA_NS.

Identification of MSTs

MSTs were identified by manual comparison with MSTs derived from commercially available pure standard compounds. Standard compounds were processed through standard addition experiments in order to obtain respective MEOX- and MSTFA-derivative mass spectra and corresponding retention time indices. Required criteria for MST identification were chromatographic co-elution within an RI window of ±2.5, and high mass spectral similarity, with matching value >700, on a scale of 0–1000. MST identification was supported by comparison with mass spectral entries from the commercial NIST02 library (National Institute of Standards and Technology). The resulting supervised and non-redundant MST library of *L. japonicus* is available as Supplementary material, NIST02 custom library file Q_LJA_ID. Identified metabolites are listed in Table 5.

Generation of a metabolite response matrix

GC-MS metabolite profiles of primary roots, lateral roots, and nodule were generated from a series of independent cultivations of fully nodulated *L. japonicus* plants. For each metabolite that was identified from the non-supervised MST library a specific mass fragment and corresponding retention-time window was selected (Table 5). The find algorithm of the MassLab software (ThermoQuest, Manchester, UK) was used to automatically retrieve peak areas from GC-MS metabolite profiles. Correct peak integration was monitored manually. Response ratios were calculated using the average response of co-analysed standard samples for normalization. All response ratios were combined into a single matrix that described the complete set of metabolite response ratios of all samples.

Principal components analysis

Principal components analysis was performed after $\log_{10}$ transformation of the above metabolite response matrix (File S3). Missing values were defined as 0 after $\log_{10}$ transformation. This substitution procedure assumed that missing values were unchanged when compared with the co-analysed standard samples. The S-Plus 2000 software package standard edition release 3 (Insightful, Berlin, Germany) was used for PCA and visualization.

Acknowledgements

We would like to thank Wolf Scheible, Alisdair Fernie, and John Lunn for helpful discussions, Peter Krüger and Dirk Steinhauser for bioinformatics support, and Peggy Lange and Grit Jochmann for technical assistance. The Max Planck Society, Deutsche Forschungs Gemeinschaft, and European Union (HPRN-CT-2000-00086) are thanked for their support of this work.

Supplementary material

The following material is available from http://www.blackwellpublishing.com/products/journals/suppmat/TPJ/TPJ2150/TPJ2150sm.htm:

Table S1 Normalized nodule and root transcript levels for all genes on the *Lotus* array

Data are from two biological replicates in both cases. A total of eight filters were hybridized to probes derived from nodule RNA (ND, four filters for each biological replicate), while five filters were hybridized to probes from roots (RT, three plus two filters for the two biological replicates). Data were normalized as described in Experimental procedures. NA indicates that little or no DNA was spotted for the corresponding clone on that particular filter (signal from reference hybridization less than 2× the local background; see Experimental procedures). GenBank accession numbers for each clone are given accept in cases where poor quality sequence was obtained; these sequences were not submitted (NS). Averages of nodule and root transcript levels, NDAVG and RTAVG, respectively, and the ratio of the two, RATIOS ND/RT are also presented. *P*-values were obtained from the Student's *t*-test

Table S2 Nodule-induced genes for cell biogenesis, cell division, intracellular transport, and protein synthesis

EST identifier and corresponding tentative consensus number from the TIGR *Lotus japonicus* Gene Index (http://www.tigr.org/tdb/tgi/ljgi/) are shown. Corresponding GenBank accession numbers are provided in Table S1. N/R denotes mean nodule/root transcript ratio. *P*-values were obtained from Student's *t*-tests.

Table S3 Nodule-repressed genes

EST identifier and corresponding tentative consensus number from the TIGR *Lotus japonicus* Gene Index (http://www.tigr.org/tdb/tgi/ljgi/) are shown. Corresponding GenBank accession numbers are provided in Table S1. N/R denotes mean nodule/root transcript ratio. *P*-values were obtained from Student's *t*-tests

File S1 The NIST02 custom library Q_LJA_NS contains all non-curated GC-MS mass spectra of MSTs from extracts of *Lotus japonicus* organs, namely nodule, primary and lateral root, flower, developing leaf and mature leaf. The spectrum name was designed to allow sorting according to plant organ, retention time index, and experiment identifier. Besides retention time index, absolute retention time, and a short sample description are given.

The file is ready to use and can be directly copied into the MSSEARCH folder of downloaded NIST98 or NIST02 mass-spectral comparison software (to be downloaded from http://www.chemdata.nist.gov/mass-spc/Srch_v1.7/index.html).

File S2 The NIST02 custom library Q_LJA_ID represents a curated and non-redundant subset of identified MSTs from the Q_LJA_NS library. The spectrum name was extended by the respective name of the metabolite derivative. Metabolites may produce more than one derivative.

The file is ready to use and can be directly copied into the MSSEARCH folder of downloaded NIST98 or NIST02 mass-spectral comparison software (to be downloaded from http://www.chemdata.nist.gov/mass-spc/Srch_v1.7/index.html).

File S3 Log_{10} transformed and fully substituted metabolite response matrix of *Lotus japonicus* root organs. Sample description and MST identification are included.

References

Arwas, R., McKay, I.A., Rowney, F.R.P., Dilworth, M.J. and Glenn, A.R. (1985) Properties of organic-acid utilization mutants of rhizobium-leguminosarum strain-300. *J. Gen. Microbiol.* **131**, 2059–2066.

Ausloos, P., Clifton, C.L., Lias, S.G., Mikaya, A.I., Stein, S.E., Tchekhovskoi, D.V., Sparkman, O.D., Zaikin, V. and Zhu, D. (1999) The critical evaluation of a comprehensive mass spectral library. *J. Am. Soc. Mass Spectrom.* **10**, 287–299.

Bardin, S., Dan, S., Osteras, M. and Finan, T.M. (1996) A phosphate transport system is required for symbiotic nitrogen fixation by *Rhizobium meliloti*. *J. Bacteriol.* **178**, 4540–4547.

Batut, J. and Boistard, P. (1994) Oxygen control in rhizobium. *Antonie Van Leeuwenhoek Int. J. Gen. Mol. Microbiol.* **66**, 129–150.

Bergersen, F.J. (1997) Regulation of nitrogen fixation in infected cells of leguminous root nodules in relation to O-2 supply. *Plant Soil*, **191**, 189–203.

Broughton, W.J. and Dilworth, M.J. (1971) Control of leghaemoglobin synthesis in snake beans. *Biochem. J.* **125**, 1075–1080.

Buschges, R., Hollricher, K., Panstruga, R. *et al.* *(1997) The barley mlo gene: a novel control element of plant pathogen resistance. Cell*, **88**, 695–705.

Christiansen, H., Hansen, A.C., Vijn, I., Pallisgaard, N., Larsen, K., Yang, W.C., Bisseling, T., Marcker, K.A. and Jensen, E.O. (1996) A novel type of DNA-binding protein interacts with a conserved sequence in an early nodulin ENOD12 promoter. *Plant Mol. Biol.* **32**, 809–821.

Colebatch, G., Kloska, S., Trevaskis, B., Freund, S., Altmann, T. and Udvardi, M.K. (2002) Novel aspects of symbiotic nitrogen fixation uncovered by transcript profiling with cDNA arrays. *Mol. Plant Microbe Interact.* **15**, 411–420.

Copeland, L., Vella, J. and Hong, Z.Q. (1989) Enzymes of carbohydrate metabolism in soybean nodules. *Phytochemistry*, **28**, 57–61.

Copeland, L., Lee, H.S. and Cowlishaw, N. (1995) Carbon metabolism in chickpea nodules. *Soil Biol. Biochem.* **27**, 381–385.

Craig, J., Barratt, P., Tatge, H. *et al.* *(1999) Mutations at the rug4 locus alter the carbon and nitrogen metabolism of pea plants through an effect on sucrose synthase. Plant J.* **17**, 353–362.

Cullimore, J.V., Gebhardt, C., Saarelainen, R., Miflin, B.J., Idler, K.B. and Barker, R.F. (1984) Glutamine synthetase of *Phaseolus vulgaris* L.: organ-specific expression of a multigene family. *J. Mol. Appl. Genet.* **2**, 589–599.

Czechowski, T., Bari, R.P., Stitt, M., Scheible, W-R. and Udvardi, M.K. (2004) Real-time RT-PCR profiling of over 1,400 *Arabidopsis* transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *Plant J.* **38**, 366–379.

Day, D.A. and Copeland, L. (1991) Carbon metabolism and compartmentation in nitrogen-fixing legume nodules. *Plant Physiol. Biochem.* **29**, 185–201.

Delauney, A.J. and Verma, D.P.S. (1990) A soybean gene encoding delta-1-pyrroline-5-carboxylate reductase was isolated by functional complementation in *Escherichia coli* and is found to be osmoregulated. *Mol. Gen. Genet.* **221**, 299–305.

Denarie, J., Debelle, F. and Prome, J.C. (1996) Rhizobium lipo-chitooligosaccharide nodulation factors: signaling molecules mediating recognition and morphogenesis. *Annu. Rev. Biochem.* **65**, 503–535.

Dormann, P. and Benning, C. (2002) Galactolipids rule in seed plants. *Trends Plant Sci.* **7**, 112–118.

Downie, J.A. and Walker, S.A. (1999) Plant responses to nodulation factors. *Curr. Opin. Plant Biol.* **2**, 483–489.

Drew, M.C. (1997) Oxygen deficiency and root metabolism: injury and acclimation under hypoxia and anoxia. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **48**, 223–250.

Endre, G., Kereszt, A., Kevei, Z., Mihacea, S., Kalo, P. and Kiss, G.B. (2002) A receptor kinase gene regulating symbiotic nodule development. *Nature*, **417**, 962–966.

Fedorova, M., van de Mortel, J., Matsumoto, P.A., Cho, J., Town, C.D., VandenBosch, K.A., Gantt, J.S. and Vance, C.P. (2002) Genome-wide identification of nodule-specific transcripts in the model legume *Medicago truncatula*. *Plant Physiol.* **130**, 519–537.

Fernie, A.R., Tiessen, A., Stitt, M., Willmitzer, L. and Geigenberger, P. (2002) Altered metabolic fluxes result from shifts in metabolite levels in sucrose phosphorylase-expressing potato tubers. *Plant Cell Environ.* **25**, 1219–1232.

Fiehn, O., Kopka, J., Trethewey, R.N. and Willmitzer, L. (2000a) Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Anal. Chem.* **72**, 3573–3580.

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R.N. and Willmitzer, L. (2000b) Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* **18**, 1157–1161.

Fischer, H.M. (1996) Environmental regulation of rhizobial symbiotic nitrogen fixation genes. *Trends Microbiol.* **4**, 317–320.

Flemetakis, E., Dimou, M., Cotzur, D., Efrose, R.C., Aivalakis, G., Colebatch, G., Udvardi, M. and Katinakis, P. (2003) A sucrose transporter, LjSUT4, is up-regulated during *Lotus japonicus* nodule development. *J. Exp. Botany*, **54**, 1789–1791.

Fortin, M.G., Morrison, N.A. and Verma, D.P.S. (1987) Nodulin-26, a peribacteroid membrane nodulin is expressed independently of

the development of the peribacteroid compartment. *Nucl. Acids Res.* **15**, 813–824.

Frugier, F., Poirier, S., Satiat-Jeunemaitre, B., Kondorosi, A. and Crespi, M. (2000) A Kruppel-like zinc finger protein is involved in nitrogen-fixing root nodule organogenesis. *Genes Dev.* **14**, 475–482.

Fujihara, S., Abe, H., Minakawa, Y., Akao, S. and Yoneyama, T. (1994) Polyamines in nodules from various plant–microbe symbiotic associations. *Plant Cell Physiol.* **35**, 1127–1134.

Gardiol, A., Arias, A., Cervenansky, C. and Martinezdrets, G. (1982) Succinate-dehydrogenase mutant of *Rhizobium meliloti*. *J. Bacteriol.* **151**, 1621–1623.

Geigenberger, P. (2003) Response of plant metabolism to too little oxygen. *Curr. Opin. Plant Biol.* **6**, 247–256.

van Ghelue, M., Ribeiro, A., Solheim, B., Akkermans, A.D., Bisseling, T. and Pawlowski, K. (1996) Sucrose synthase and enolase expression in actinorhizal nodules of *Alnus glutinosa*: comparison with legume nodules. *Mol. Gen. Genet.* **250**, 437–446.

Gordon, A.J. and James, C.L. (1997) Enzymes of carbohydrate and amino acid metabolism in developing and mature nodules of white clover. *J. Exp. Botany*, **48**, 895–903.

Gordon, A.J., Minchin, F.R., James, C.L. and Komina, O. (1999) Sucrose synthase in legume nodules is essential for nitrogen fixation. *Plant Physiol.* **120**, 867–878.

Gregerson, R.G., Miller, S.S., Petrowski, M., Gantt, J.S. and Vance, C.P. (1994) Genomic structure, expression and evolution of the alfalfa aspartate aminotransferase genes. *Plant Mol. Biol.* **25**, 387–399.

Guenther, J.F. and Roberts, D.M. (2000) Water-selective and multifunctional aquaporins from *Lotus japonicus* nodules. *Planta*, **210**, 741–748.

Hata, S., Izui, K. and Kouchi, H. (1998) Expression of a soybean nodule-enhanced phosphoenolpyruvate carboxylase gene that shows striking similarity to another gene for a house-keeping isoform. *Plant J.* **13**, 267–273.

Heard, J. and Dunn, K. (1995) Symbiotic induction of a MADS-Box gene during development of alfalfa root-nodules. *Proc. Natl Acad. Sci. USA*, **92**, 5273–5277.

Imsande, J., Berkemeyer, M., Scheibe, R., Schumann, U., Gietl, C. and Palmer, R.G. (2001) A soybean plastid-targeted NADH-malate dehydrogenase: cloning and expression analyses. *Am. J. Botany*, **88**, 2136–2142.

Izmailov, S.F., Andreeva, I.N. and Kozharinova, G.M. (1999) Subcellular calcium localization in the root nodules of legumes. *Russ. J. Plant Physiol.* **46**, 93–101.

Journet, E.P., van Tuinen, D., Gouzy, J. *et al.* *(2002) Exploring root symbiotic programs in the model legume Medicago truncatula* using EST analysis. *Nucl. Acids Res.* **30**, 5579–5592.

Kahn, M.L., McDermott, T.R. and Udvardi, M.K. (1998) Carbon and nitrogen metabolism in the Rhizobiaceae. In *The Rhizobiaceae* (**Hooykaas, P.J.J.**, ed.). Dordrecht: Kluwer Academic Publishers, pp. 461–465.

Kaiser, B.N., Finnegan, P.M., Tyerman, S.D., Whitehead, L.F., Bergersen, F.J., Day, D.A. and Udvardi, M.K. (1998) Characterization of an ammonium transport protein from the peribacteroid membrane of soybean nodules. *Science*, **281**, 1202–1206.

Kavroulakis, N., Flemetakis, E., Aivalakis, G. and Katinakis, P. (2000) Carbon metabolism in developing soybean root nodules: the role of carbonic anhydrase. *Mol. Plant Microbe Interact.* **13**, 14–22.

Klok, E.J., Wilson, I.W., Wilson, D., Chapman, S.C., Ewing, R.M., Somerville, S.C., Peacock, W.J., Dolferus, R. and Dennis, E.S. (2002) Expression profile analysis of the low-oxygen response in *Arabidopsis* root cultures. *Plant Cell*, **14**, 2481–2494.

Komina, O., Zhou, Y., Sarath, G. and Chollet, R. (2002) In vivo and in vitro phosphorylation of membrane and soluble forms of soybean nodule sucrose synthase. *Plant Physiol.* **129**, 1664–1673.

Kouchi, H. and Hata, S. (1993) Isolation and characterization of novel nodulin cDNAs representing genes expressed at early stages of soybean nodule development. *Mol. Gen. Genet.* **238**, 106–119.

Krusell, L., Madsen, L.H., Sato, S. *et al.* *(2002) Shoot control of root development and nodulation is mediated by a receptor-like kinase. Nature*, **420**, 422–426.

Lara, M., Cullimore, J.V., Lea, P.J., Miflin, B.J., Johnston, A.W.B. and Lamb, J.W. (1983) Appearance of a novel form of plant glutamine-synthetase during nodule development in *Phaseolus vulgaris* L. *Planta*, **157**, 254–258.

Lee, J.W., Zhang, Y.X., Weaver, C.D., Shomer, N.H., Louis, C.F. and Roberts, D.M. (1995) Phosphorylation of nodulin-26 on serine-262 affects its voltage-sensitive channel activity in planar lipid bilayers. *J. Biol. Chem.* **270**, 27051–27057.

Liu, J.P. and Zhu, J.K. (1997) Proline accumulation and salt-stress-induced gene expression in a salt-hypersensitive mutant of *Arabidopsis*. *Plant Physiol.* **114**, 591–596.

Lodwig, E.M., Hosie, A.H.F., Bordes, A., Findlay, K., Allaway, D., Karunakaran, R., Downie, J.A. and Poole, P.S. (2003) Amino-acid cycling drives nitrogen fixation in the legume – *Rhizobium symbiosis. Nature*, **422**, 722–726.

Long, S.R. (2001) Genes and signals in the *Rhizobium*-legume symbiosis. *Plant Physiol.* **125**, 69–72.

Miller, S.S., Driscoll, B.T., Gregerson, R.G., Gantt, J.S. and Vance, C.P. (1998) Alfalfa malate dehydrogenase (MDH): molecular cloning and characterization of five different forms reveals a unique nodule-enhanced MDH. *Plant J.* **15**, 173–184.

Mimura, T. (1999) Regulation of phosphate transport and homeostasis in plant cells. *Int. Rev. Cytol.* **191**, 141–200.

Mithofer, A. (2002) Suppression of plant defence in rhizobia-legume symbiosis. *Trends Plant Sci.* **7**, 440–444.

Moreau, S., Thomson, R.M., Kaiser, B.N., Trevaskis, B., Guerinot, M.L., Udvardi, M.K., Puppo, A. and Day, D.A. (2002) GmZIP1 encodes a symbiosis-specific zinc transporter in soybean. *J. Biol. Chem.* **277**, 4738–4746.

Nakagawa, T., Izumi, T., Banba, M., Umehara, Y., Kouchi, H., Izui, K. and Hata, S. (2003) Characterization and expression analysis of genes encoding phosphoenolpyruvate carboxylase and phosphoenolpyruvate carboxylase kinase of *Lotus japonicus*, a model legume. *Mol. Plant Microbe Interact.* **16**, 281–288.

Newton, W.E. (2000) *Nitrogen Fixation: From Molecules to Crop Productivity*. Dordrecht: Kluwer, pp. 3–8.

Nishimura, R., Hayashi, M., Wu, G.J. *et al.* *(2002) HAR1 mediates systemic regulation of symbiotic organ development. Nature*, **420**, 426–429.

Parsons, R. and Day, D.A. (1990) Mechanism of soybean nodule adaptation to different oxygen pressures. *Plant Cell Environ.* **13**, 501–512.

Pathirana, S.M., Vance, C.P., Miller, S.S. and Gantt, J.S. (1992) Alfalfa root nodule phosphoenolpyruvate carboxylase: characterization of the cDNA and expression in effective and plant-controlled ineffective nodules. *Plant Mol. Biol.* **20**, 437–450.

de la Pena, T.C., Frugier, F., McKhann, H.I., Bauer, P., Brown, S., Kondorosi, A. and Crespi, M. (1997) A carbonic anhydrase gene is induced in the nodule primordium and its cell-specific expression is controlled by the presence of *Rhizobium* during development. *Plant J.* **11**, 407–420.

Penheiter, A.R., Duff, S.M.G. and Sarath, G. (1997) Soybean root nodule acid phosphatase. *Plant Physiol.* **114**, 597–604.

Perlick, A.M. and Puhler, A. (1993) A survey of transcripts expressed specifically in root nodules of broadbean (*Vicia faba* L.). *Plant Mol. Biol.* **22**, 957–970.
Raghothama, K.G. (1999) Phosphate acquisition. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **50**, 665–693.
Reddy, A.S.N. (2001) Calcium: silver bullet in signaling. *Plant Sci.* **160**, 381–404.
Reynolds, P.H., Smith, L.A., Dickson, J.M., Jones, W.T., Jones, S.D., Rodber, K.A., Carne, A. and Liddane, C.P. (1992) Molecular cloning of a cDNA encoding aspartate aminotransferase-P2 from lupin root nodules. *Plant Mol. Biol.* **19**, 465–472.
Riechmann, J.L., Heard, J., Martin, G. *et al.* *(2000) Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.
Ritte, G., Lloyd, J.R., Eckermann, N., Rottmann, A., Kossmann, J. and Steup, M. (2002) The starch-related R1 protein is an alpha-glucan, water dikinase. *Proc. Natl Acad. Sci. USA*, **99**, 7166–7171.
Rivers, R.L., Dean, R.M., Chandy, G., Hall, J.E., Roberts, D.M. and Zeidel, M.L. (1997) Functional analysis of nodulin 26, an aquaporin in soybean root nodule symbiosomes. *J. Biol. Chem.* **272**, 16256–16261.
Roessner, U., Wagner, C., Kopka, J., Trethewey, R.N. and Willmitzer, L. (2000) Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J.* **23**, 131–142.
Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L. and Fernie, A. (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell*, **13**, 11–29.
Ronson, C.W., Lyttleton, P. and Robertson, J.G. (1981) C4-dicarboxylate transport mutants of *Rhizobium trifolii* form ineffective nodules on *Trifolium repens*. *Proc. Natl Acad. Sci. USA-Biol. Sci.* **78**, 4284–4288.
Sachs, M.M., Subbaiah, C.C. and Saab, I.N. (1996) Anaerobic gene expression and flooding tolerance in maize. *J. Exp. Botany*, **47**, 1–15.
Schauser, L., Roussis, A., Stiller, J. and Stougaard, J. (1999) A plant regulator controlling development of symbiotic root nodules. *Nature*, **402**, 191–195.
Schuller, K.A. and Werner, D. (1993) Phosphorylation of soybean (*Glycine max* L.) nodule phosphoenolpyruvate carboxylase in vitro decreases sensitivity to inhibition by L-malate. *Plant Physiol.* **101**, 1267–1273.
Searle, I.R., Men, A.E., Laniya, T.S., Buzas, D.M., Iturbe-Ormaetxe, I., Carroll, B.J. and Gresshoff, P.M. (2003) Long-distance signaling in nodulation directed by a CLAVATA1-like receptor kinase. *Science*, **299**, 109–112.
Shi, L., Twary, S.N., Yoshioka, H., Gregerson, R.G., Miller, S.S., Samac, D.A., Gantt, J.S., Unkefer, P.J. and Vance, C.P. (1997) Nitrogen assimilation in alfalfa: isolation and characterization of an asparagine synthetase gene showing enhanced expression in root nodules and dark-adapted leaves. *Plant Cell*, **9**, 1339–1356.
Snedden, W.A. and Fromm, H. (1998) Calmodulin, calmodulin-related proteins and plant responses to the environment. *Trends Plant Sci.* **3**, 299–304.
Stacey, G., Koh, S., Granger, C. and Becker, J.M. (2002) Peptide transport in plants. *Trends Plant Sci.* **7**, 257–263.
Stein, S.E. (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *J. Am. Soc. Mass Spectrom.* **10**, 770–781.
Stougaard, J. (2000) Regulators and regulation of legume root nodule development. *Plant Physiol.* **124**, 531–540.
Stougaard, J. (2001) Genetics and genomics of root symbiosis. *Curr. Opin. Plant Biol.* **4**, 328–335.
Stracke, S., Kistner, C., Yoshida, S. *et al.* *(2002) A plant receptor-like kinase required for both bacterial and fungal symbiosis. Nature*, **417**, 959–962.
Streeter, J.G. (1995) Recent developments in carbon transport and metabolism in symbiotic systems. *Symbiosis*, **19**, 175–196.
Suganuma, N., Okada, Y. and Kanayama, Y. (1997) Isolation of a cDNA for nodule-enhanced phosphoenolpyruvate carboxylase from pea and its expression in effective and plant-determined ineffective pea nodules. *J. Exp. Botony*, **48**, 1165–1173.
Swaraj, K. and Bishnoi, N.R. (1999) Effect of salt stress on nodulation and nitrogen fixation in legumes. *Indian J. Exp. Biol.* **37**, 843–848.
Szczyglowski, K., Kapranov, P., Hamburger, D. and de Bruijn, F.J. (1998) The *Lotus japonicus* LjNOD70 nodulin gene encodes a protein with similarities to transporters. *Plant Mol. Biol.* **37**, 651–661.
Tansengco, M.L., Hayashi, M., Kawaguchi, M., Imaizumi-Anraku, H. and Murooka, Y. (2003) Crinkle, a novel symbiotic mutant that affects the infection thread growth and alters the root hair, trichome, and seed development in *Lotus japonicus*. *Plant Physiol.* **131**, 1054–1063.
Thummler, F. and Verma, D.P. (1987) Nodulin-100 of soybean is the subunit of sucrose synthase regulated by the availability of free heme in nodules. *J. Biol. Chem.* **262**, 14730–14736.
Tjepkema, J.D. and Yocum, C.S. (1974) Measurement of oxygen partial-pressure within soybean nodules by oxygen microelectrodes. *Planta*, **119**, 351–360.
Tsai, F.Y. and Coruzzi, G.M. (1990) Dark-induced and organ-specific expression of two asparagine synthetase genes in *Pisum sativum*. *EMBO J.* **9**, 323–332.
Udvardi, M.K. and Day, D.A. (1997) Metabolite transport across symbiotic membranes of legume nodules. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **48**, 493–523.
Udvardi, M.K. and Kahn, M.L. (1993) Evolution of *Rhizobium* legume symbiosis: why do bacteroids fix nitrogen? *Symbiosis* **14**, 87–101.
Udvardi, M.K., Price, G.D., Gresshoff, P.M. and Day, D.A. (1988) A dicarboxylate transporter on the peribacteroid membrane of soybean nodules. *FEBS Lett.* **231**, 36–40.
Uhde-Stone, C., Zinn, K.E., Ramirez-Yanez, M., Li, A.G., Vance, C.P. and Allan, D.L. (2003) Nylon filter arrays reveal differential gene expression in proteoid roots of white lupin in response to phosphorus deficiency. *Plant Physiol.* **131**, 1064–1079.
Vance, C.P. and Johnson, L.E.B. (1983) Plant determined ineffective nodules in alfalfa (*Medicago sativa*) – structural and biochemical comparisons. *Can. J. Botany-Revue Canadienne De Botanique*, **61**, 93–106.
Vance, C.P., Reibach, P.H. and Pankhurst, C.E. (1987) Symbiotic properties of *Lotus pedunculatus* root nodules induced by *Rhizobium loti* and *Bradyrhizobium* sp. (Lotus). *Physiol. Plantarum*, **69**, 435–442.
Vance, C.P., Gregerson, R.G., Robinson, D.L., Miller, S.S. and Gantt, J.S. (1994) Primary assimilation of nitrogen in alfalfa nodules: molecular features of the enzymes involved. *Plant Sci.* **101**, 51–64.
Wadham, C., Winter, H. and Schuller, K.A. (1996) Regulation of soybean nodule phosphoenolpyruvate carboxylase *in vivo*. *Physiol. Plantarum*, **97**, 531–535.
Wagner, C., Sefkow, M. and Kopka, J. (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry*, **62**, 887–900.
Wandrey, M., Trevaskis, B., Brewin, N. and Udvardi, M.K. (2004) Molecular and cell biology of a family of VDAC porins in *Lotus japonicus*. *Plant Physiol.* **13**, 182–193.

Weaver, C.D. and Roberts, D.M. (1991) Phosphorylation of nodulin-26 by a calcium-dependent protein-kinase. *FASEB J.* **5**, A426–A426.

Weaver, C.D., Crombie, B., Stacey, G. and Roberts, D.M. (1991) Calcium-dependent phosphorylation of symbiosome membrane-proteins from nitrogen-fixing soybean nodules – evidence for phosphorylation of nodulin-26. *Plant Physiol.* **95**, 222–227.

Weaver, C.D., Shomer, N.H., Louis, C.F. and Roberts, D.M. (1994) Nodulin-26, a nodule-specific symbiosome membrane-protein from soybean, is an ion-channel. *J. Biol. Chem.* **269**, 17858–17862.

Xu, W.X., Zhou, Y. and Chollet, R. (2003) Identification and expression of a soybean nodule-enhanced PEP-carboxylase kinase gene (NE-PpcK) that shows striking up-/down-regulation in vivo. *Plant J.* **34**, 441–452.

Yu, T.S., Kofler, H., Hausler, R.E. *et al.* *(2001) The Arabidopsis* sex1 mutant is defective in the R1 protein, a general regulator of starch degradation in plants, and not in the chloroplast hexose transporter. *Plant Cell*, **13**, 1907–1918.

Zhang, X.Q. and Chollet, R. (1997a) A Ca^{2+}-independent protein kinase is involved in phosphorylation of phospho enol pyruvate carboxylase in soybean root nodules and is up/down-regulated by photosynthate supply from the shoots. *Plant Physiol.* (Suppl. 114), 33.

Zhang, X.Q. and Chollet, R. (1997b) Phosphoenolpyruvate carboxylase protein kinase from soybean root nodules: partial purification, characterization, and up/down-regulation by photosynthate supply from the shoots. *Arch. Biochem. Biophys.* **343**, 260–268.

Zhang, X.Q. and Chollet, R. (1997c) Seryl-phosphorylation of soybean nodule sucrose synthase (nodulin-100) by a Ca^{2+}-dependent protein kinase. *FEBS Lett.* **410**, 126–130.

Zielinski, R.E. (1998) Calmodulin and calmodulin-binding proteins in plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **49**, 697–725.

Zucchero, J.C., Caspi, M. and Dunn, K. (2001) ngl9: a third MADS box gene expressed in alfalfa root nodules. *Mol. Plant Microbe Interact.* **14**, 1463–1467.

© Blackwell Publishing Ltd, *The Plant Journal*, (2004), **39**, 487–512

Lotus japonicus Metabolic Profiling. Development of Gas Chromatography-Mass Spectrometry Resources for the Study of Plant-Microbe Interactions

Guilhem G. Desbrosses[1], Joachim Kopka, and Michael K. Udvardi*

Max Planck Institute of Molecular Plant Physiology, 14476 Golm, Germany

Symbiotic nitrogen fixation (SNF) in legume root nodules requires differentiation and integration of both plant and bacterial metabolism. Classical approaches of biochemistry, molecular biology, and genetics have revealed many aspects of primary metabolism in legume nodules that underpin SNF. Functional genomics approaches, especially transcriptomics and proteomics, are beginning to provide a more holistic picture of the metabolic potential of nodules in model legumes like *Medicago truncatula* and *Lotus japonicus*. To extend these approaches, we have established protocols for nonbiased measurement and analysis of hundreds of metabolites from *L. japonicus*, using gas chromatography coupled with mass spectrometry. Following creation of mass spectral tag libraries, which represent both known and unknown metabolites, we measured and compared relative metabolite levels in nodules, roots, leaves, and flowers of symbiotic plants. Principal component analysis of the data revealed distinct metabolic phenotypes for the different organs and led to the identification of marker metabolites for each. Metabolites that were enriched in nodules included: octadecanoic acid, asparagine, glutamate, homoserine, cysteine, putrescine, mannitol, threonic acid, gluconic acid, glyceric acid-3-P, and glycerol-3-P. Hierarchical cluster analysis enabled discrimination of 10 groups of metabolites, based on distribution patterns in diverse Lotus organs. The resources and tools described here, together with ongoing efforts in the areas of genome sequencing, and transcriptome and proteome analysis of *L. japonicus* and *Mesorhizobium loti*, should lead to a better understanding of nodule metabolism that underpins SNF.

The legume family comprises approximately 700 genera with more than 18,000 species, which occupy niches in almost every environment on earth (Polhill et al., 1981; Doyle and Luckow, 2003). A key to the success of this family was the evolution of mutualistic symbioses with bacteria of the family Rhizobiaceae, which enabled early legumes to utilize atmospheric N_2 as a source of nitrogen, especially when colonizing soils lacking mineral or organic nitrogen. Today, symbiotic nitrogen fixation (SNF) by rhizobia in legumes takes place in specialized plant organs called nodules. Nodules develop from cortical cells of the root or stem after contact with rhizobia in the soil (Brewin, 1991). Mature, nitrogen-fixing nodules consist of several layers of uninfected plant cells surrounding a central zone of infected and noninfected plants cells. Infected plant cells typically contain thousands of differentiated, nitrogen-fixing rhizobia, called bacteroids, which are separated from the cytoplasm, either individually or in small groups, by a unique plant membrane called the peribacteroid or symbiosome membrane (SM; Roth et al., 1988; Udvardi and Day, 1997). Microaerobic conditions within legume nodules result from restricted oxygen influx across the outer cell layers of nodules, binding and transport of oxygen by leghemoglobin in the cytoplasm of plant cells, and high rates of respiration by bacteroids and mitochondria in these cells (Appleby, 1984). Low steady-state oxygen concentrations within nodules (in the nanomolar range) have profound effects on plant and bacterial metabolism in nodules. For instance, microaerobiosis is a prerequisite for activity of the oxygen-labile bacteroid enzyme, nitrogenase (Robson and Postgate, 1980).

SNF involves the mutually beneficial exchange of reduced carbon from the plant for reduced nitrogen from the bacteria (Udvardi and Day, 1997), which requires metabolic differentiation of both organisms. Suc, delivered via the phloem, is the primary source of carbon and energy for nodule metabolism (Gordon et al., 1999). However, genetic studies with rhizobial mutants, together with biochemical studies of metabolite transport across the SM and bacteroid membranes, indicate that dicarboxylic acids, especially malate, rather than sugars, are the main source of carbon supplied to bacteroids for SNF (Ronson et al., 1981; Gardiol et al., 1987; Udvardi et al., 1988). An important aspect of plant differentiation during nodule development is the induction of genes and proteins that convert sugars to malate via glycolysis and carbon fixation (Pathirana et al., 1992; Miller et al., 1998; Colebatch et al., 2002, 2004). At about the same time, decreasing oxygen within nodules triggers induction of rhizobial genes for nitrogenase and high-affinity oxidases, which enable bacteroid nitrogen fixation and respiration under these conditions (Batut and Boistard, 1994; Fischer, 1996; Sciotti et al., 2003).

[1] Present address: Université Montpellier 2, CC 002, Place Eugène Bataillon, F–34095 Montpellier cedex 05, France.

* Corresponding author; e-mail udvardi@mpimp-golm.mpg.de; fax 49–331–567–8250.

Article, publication date, and citation information can be found at www.plantphysiol.org/cgi/doi/10.1104/pp.104.054957.

Finally, induction of plant genes for ammonium assimilation facilitates rapid incorporation of nitrogen into amino acids and other nitrogen compounds for export to the rest of the plant (Vance et al., 1994; Colebatch et al., 2004). These are some of the principal metabolic changes that occur during nodule development and differentiation, and most studies of nodule metabolism have focused on one or more of these aspects in a variety of different legumes. Few studies have attempted to look more broadly at nodule metabolism in a single, model species. To facilitate such studies, we have developed resources for transcriptome (Colebatch et al., 2002, 2004) and metabolome analyses in the model legume, *Lotus japonicus*.

In the past, most studies on legume metabolites analyzed a few compounds from preselected classes such as sugars, amino and organic acids, thiols, saponins, and phenolics, using a range of instrumentation, including HPLC (Streeter, 1987; Matamoros et al., 1999; Chen et al., 2003), thin layer chromatography (Khalil and Eladawy, 1994; Steele et al., 1999), and gas chromatography (GC; Streeter and Bosler, 1976; Streeter, 1980; Karoutis et al., 1992). Such studies can best be described as targeted metabolite analysis (Fiehn, 2002), where analysis concentrates on a few, well-defined metabolites, which are often part of well-characterized metabolic pathways. Relatively little attention has been focused on unknown compounds representing potentially novel metabolism. Recent development or refinement of a variety of analytical platforms, including GC-mass spectrometry (GC-MS; Fiehn et al., 2000; Roessner et al., 2000; Wagner et al., 2003) and liquid chromatography-MS (Huhman and Sumner, 2002; Tolstikov and Fiehn, 2002; Chen et al., 2003; Tolstikov et al., 2003), together with software developments, such as GC-MS chromatogram alignment tools (Duran et al., 2003) and deconvolution programs to extricate MS data from overlapping chromatographic peaks (Stein, 1999), have enabled high-throughput, nonbiased analysis of thousands of metabolites from plants and other organisms. These tools afford not only a much broader view of metabolites and metabolism but also the opportunity to discover novel metabolites and previously unknown aspects of metabolism (Fiehn et al., 2000; Sumner et al., 2003). Here, we describe the use of GC-MS to characterize the metabolome of the model legume, *L. japonicus*. Following the creation of mass spectral tag (MST) libraries, which represent both known and unknown metabolites, we measured and compared relative metabolite levels in nodules, roots, leaves, and flowers of symbiotic plants. Principal component analysis (PCA) and hierarchical cluster analysis (HCA), revealed discrete metabolic phenotypes for the different organs and led to the identification of metabolite markers for each. A large number of novel metabolites, including 2-methylcitrate and many still unidentified metabolites, were uncovered, which alludes to previously unknown aspects of metabolism in nodules and other organs.

RESULTS

GC-MS Chromatograms of *L. japonicus* Organs and Establishment of MST Libraries

Gas chromatograms of nodules, lateral and primary roots, developing and mature leaves, and flowers from *L. japonicus* plants, harvested 12 weeks after germination and inoculation with *Mesorhizobium loti* strain R7A, revealed reproducible and organ-specific features (Fig 1). About 40 major polar metabolite derivatives were detectable by eye from the GC traces, together with a multitude of minor constituents, which are barely or not at all visible on the scale shown in Figure 1.

GC separates complex mixtures of metabolite derivatives into a series of compounds that enter the mass spectrometer and are subsequently ionized, fragmented, and detected. Each metabolite is, therefore, represented by one or more ionic fragments of precise mass, which together can serve as a tag for that metabolite. We have termed these MST, by analogy to expressed sequence tags of genes. Each MST has properties that facilitate unequivocal identification of the parent metabolite, following comparison to the pure reference compound (Wagner et al., 2003). The properties of an MST are: (1) gas chromatographic retention, which is best characterized by a retention time index (RI), and (2) a specific composition of fragments, which are each characterized by a mass-to-charge ratio (*m/z*). A library of MSTs was derived from a set of *L. japonicus* organs using the automated mass spectral deconvolution and identification system, AMDIS (Stein, 1999).

Mass fragments that belong to one MST have the same RI and occur in fixed relative abundance, independent of metabolite concentration. Therefore, any single fragment or set of fragments with identical RI can be used for the quantification of metabolites. As a rule, choice of mass fragments for quantitative purposes must be selective, i.e. only those fragments that are unique to an MST can be used. Mass fragments that are common to coeluting MSTs, i.e. fragments with similar RIs and identical *m/z*, must be avoided for quantification purposes.

In this work, fragments used for metabolite quantification were identified by *m/z*, RI, and name of MST to which the fragment belongs. If the MST represents a known or identified metabolite, we add the name of the respective metabolite derivative. We used the following nomenclature: *m/z* of the selected GC-MS fragments followed by RI and MST name both separated by the underline character; for example, mass fragment 292_2014_glucaric acid (6TMS) or 333_2014_glucaric acid (6TMS; e.g. see Fig. 4). MSTs that remain unidentified were classified tentatively by best matching mass spectra from a custom and a commercial NIST02 library (Institute of Standards and Technology, Gaithersburg, MD). A tentative match required a score >600 on a scale of 0 to 1,000. To

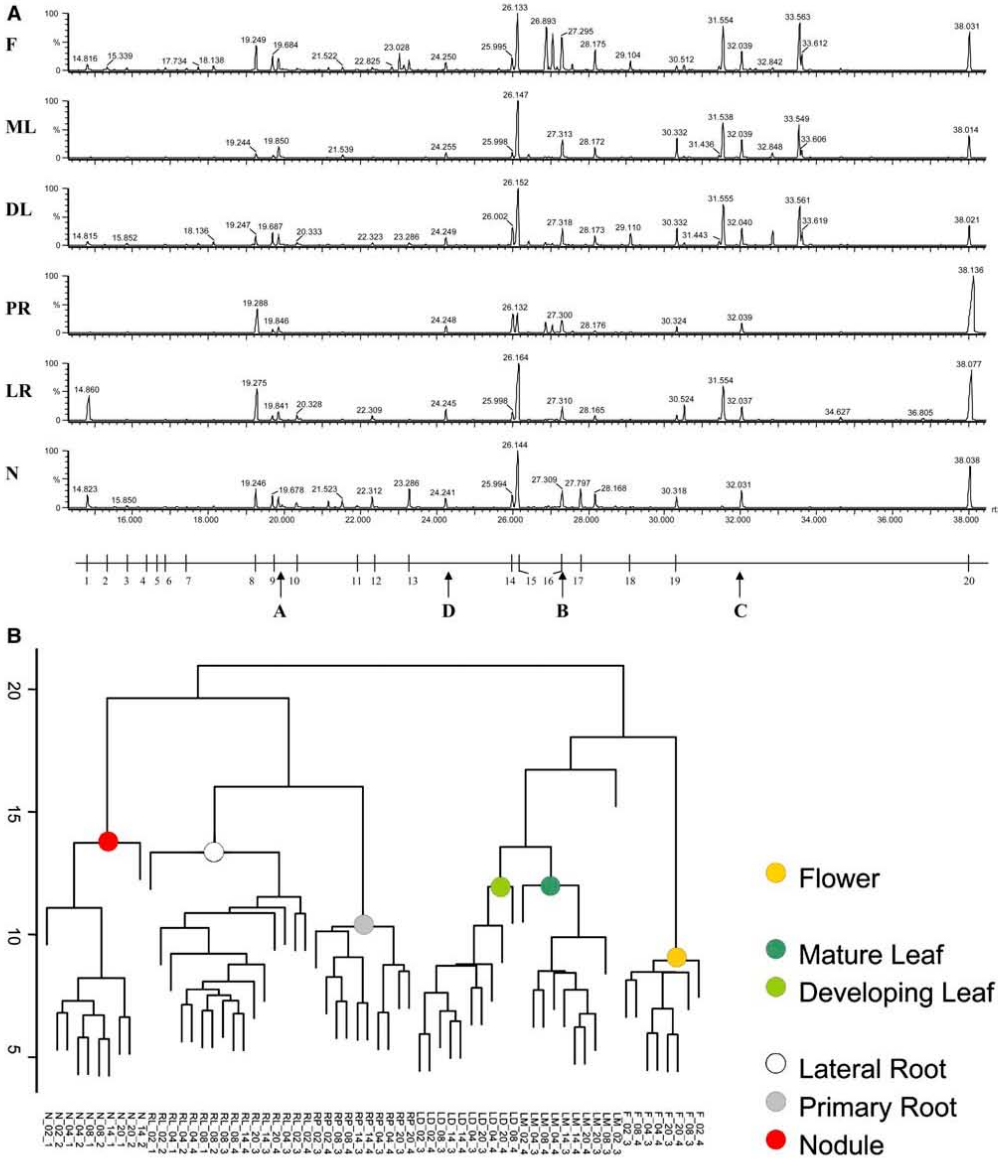


Figure 1. Typical GC-MS profiles (A) and hierarchical cluster analysis (B) of polar extracts from different organs of *L. japonicus*. Profiles were generated from nodules (N), lateral roots (LR), primary roots (PR), developing leaves (DL), mature leaves (ML), and flowers (F). Typical major MSTs represent: (1) phosphate, (2) Pro, (3) succinate, (4) glycerate, (5) fumarate, (6) Ser, (7) threonate, (8) citramalate, (9) malate, (10) Asp, (11) Asn (MST with 4 trimethylsilyl groups), (12) Gln, (13) Asn (MST with 3 trimethylsilyl groups), (14) citrate, (15) pinitol, (16) Fru, (17) ononitol, (18) saccharic acid, (19) myoinositol, and (20) Suc. Arrows indicate internal standard substances, (A) *n*-pentadecane, (B) *n*-nonadecane, (C) *n*-docosane, and (D) ribitol.

Table I. *Identified and unidentified metabolites of L. japonicus organs, GC-MS characteristics (fragment mass, RI), influence in principal components of Figure 2, differential distribution, and metabolite class membership as shown in Figure 6.*

Differences in metabolite levels that were significant at $P \leq 0.01$ are indicated in bold format.

	m/z	RI, Median	RI, SD	Influence in PCA Component (Metabolite Lists among the Top 25 of Loadings Values)	Response Ratio (Nodule/ Plant)	Response Ratio (Nodule/ Root)	Response Ratio (Root, Nodule/ Shoot)	Response Ratio (Root/ Leaf)	Response Ratio (Lateral Root, Primary Root)	Response Ratio (Developing Leaf, Mature Leaf)	Response Ratio (Flower/ Other Organs)	Cluster Membership[c]
Amino Acids												
Gly	248	1,313	1.4		2.6	3.3	1.1	1.4	2.8	**6.5**	2.3	2
L-Asn	116	1,686	3.3	2, 3	**17.6**	**21.4**	**9.1**	2.3	0.6	47.7	0.5	2
L-Homoserine	218	1,455	1.9		**5.9**	**4.9**	**3.5**	2.1	0.2	4.4	0.7	2
L-Glu	246	1,633	2.9	2	**5.2**	**11.8**	1.4	**0.4**	**0.2**	2.5	1.3	2
L-Cys	220	1,561	3.2		**5.6**	**21.0**	1.1	**0.2**	0.9	0.2	1.0	2
L-Ala	116	1,095	3.1		1.5	**2.7**	**0.6**	**0.5**	2.0	1.7	**1.8**	5
Pyroglutamic acid, L-Gln, L-Glu[a]	258	1,528	1.9		1.4	**2.2**	0.8	0.6	**0.2**	4.2	1.4	5
2-Aminoadipic acid	260	1,728	1.7		2.0	**18.8**	0.9	1.8	0.5	18.4	2.1	5
L-Val	144	1,221	1.7		0.9	1.9	0.5	0.8	1.1	0.7	**5.3**	6
L-Tyr	218	1,941	2.0		0.9	**4.5**	0.3	**0.4**	1.7	1.5	**9.7**	6
L-Met	176	1,523	2.5		1.0	**3.0**	0.5	0.9	0.3	2.9	**6.7**	6
L-Gln	156	1,786	3.3	1, 2, 3	1.4	**291.0**	0.3	0.1	1.0	12.7	**14.6**	6
L-Phe	192	1,637	2.8		1.1	**2.6**	0.5	1.1	**0.1**	1.7	**6.1**	6
L-Leu	232	1,279	1.8	3	0.4	1.6	0.2	**3.2**	0.3	0.7	**19.5**	6
L-Pro	142	1,304	1.4	1	0.1	**6.2**	**0.0**	**0.1**	0.5	3.7	**27.7**	6
L-Trp	202	2,217	3.9	2	0.0	12.3	0.0	0.2	0.0	-	**67.2**	6
L-Ser	204	1,371	1.2		0.5	**3.2**	**0.1**	**0.1**	**0.5**	1.2	**5.7**	6
L-Thr	219	1,395	1.3		0.4	1.3	**0.2**	**0.3**	**0.6**	**2.5**	**4.7**	6
β-Ala	174	1,432	0.8		**0.3**	**0.3**	0.8	**1.9**	**0.7**	**2.7**	**3.0**	6
L-Orn, L-Arg, L-Citrulline[a]	142	1,822	2.0	3, 4	0.4	0.6	0.4	4.4	0.0	12.8	**8.3**	6
L-Lys	156	1,922	2.9		1.6	**7.1**	0.7	0.9	0.1	4.3	**3.9**	6
L-Ile	158	1,302	1.9		0.7	3.3	**0.2**	**0.1**	**0.4**	**4.3**	**2.7**	7
L-Asp	232	1,526	1.8	4	0.9	1.0	0.9	1.4	**0.1**	3.0	**2.1**	9
4-Aminobutyric acid	304	1,531	1.4		0.6	0.8	**0.6**	0.7	**0.3**	1.0	1.5	9
Organic Acids												
Lactic acid[b]	219	1,049	1.3		1.0	0.8	**1.5**	**1.6**	1.7	0.9	0.8	1
Octadecanoic acid	341	2,247	2.9	2	**52.1**	**110.9**	14.1	0.6	0.1	1.2	0.3	2
2,3,4-Trihydroxybutyric acid (threonic acid)	292	1,570	1.2		**6.7**	**18.4**	1.6	**0.4**	2.1	1.5	1.5	2
Gluconic acid	333	2,003	1.8		**4.4**	3.7	2.1	3.2	4.4	1.2	1.4	2
2-Ketoglutaric acid	198	1,593	1.9		1.9	**3.4**	0.8	**0.3**	2.0	1.9	**0.2**	3
DL-2-methylcitric acid	287	1,842	0.4		**2.3**	**4.4**	0.7	**0.5**	**0.5**	0.8	**2.1**	5
Hexadecanoic acid	313	2,052	1.1		1.4	**1.6**	0.9	0.9	1.0	1.0	**1.5**	5
Glucuronic acid	160	1,938	1.2		0.2	**1.9**	**0.1**	**0.2**	**0.8**	1.1	**13.8**	6
Malic acid	335	1,493	1.8		0.9	**1.8**	**0.5**	**0.5**	1.1	1.4	**2.3**	6
Glutaric acid	158	1,416	1.7		1.7	**4.3**	**0.5**	**0.4**	1.7	2.4	**3.5**	6
Shikimic acid	204	1,822	0.9		0.8	1.8	**0.4**	**0.3**	1.0	**2.3**	2.1	7
Maleic acid	245	1,313	2.3		1.1	1.5	0.8	0.6	0.0	0.4	0.9	8
2,3,4-Trihydroxybutyric acid (erythronic acid)	292	1,550	1.3	1	**0.1**	**1.6**	**0.0**	**0.0**	0.6	**0.5**	1.6	8
Succinic acid	147	1,318	2.0		0.2	**0.4**	**0.2**	**0.3**	1.7	**0.4**	**3.1**	8
Fumaric acid	245	1,363	0.9		0.6	**2.8**	**0.2**	**0.2**	**0.9**	**0.3**	**3.9**	8
D-(-)-Quinic acid	345	1,862	1.1		0.3	**2.0**	**0.1**	**0.1**	1.3	1.1	**4.0**	8
Gulonic acid	333	1,965	1.4		0.2	1.1	**0.1**	**0.1**	1.2	0.8	**3.5**	8
cis-Aconitic acid	229	1,763	1.7		0.5	1.1	**0.3**	**0.4**	0.9	1.2	**2.2**	8
Glucaric acid	333	2,014	0.9	1	**0.1**	1.3	**0.0**	**0.0**	**0.6**	1.2	**2.8**	8
Citramalic acid	349	1,474	0.5		**0.3**	**0.2**	1.4	**2.5**	**1.0**	0.5	1.6	9
Citric acid	375	1,829	0.8		**0.2**	**0.2**	0.6	0.8	**2.4**	2.1	1.2	10
Isocitric acid	245	1,832	2.2		**0.2**	**0.2**	0.7	1.0	–	3.2	1.6	10
Galactonic acid	333	1,999	1.4		0.7	1.8	**0.3**	**0.2**	**0.2**	**0.3**	1.2	11

(*Table continues on following page.*)

Table I. *(Continued from previous page.)*

	m/z	RI, Median	RI, SD	Influence in PCA Component (Metabolite Lists among the Top 25 of Loadings Values)	Response Ratio (Nodule /Plant)	Response Ratio (Nodule /Root)	Response Ratio (Root, Nodule/ Shoot)	Response Ratio (Root/ Leaf)	Response Ratio (Lateral Root, Primary Root)	Response Ratio (Developing Leaf, Mature Leaf)	Response Ratio (Flower/ Other Organs)	Cluster Membership[c]
Glyceric acid	292	1,341	2.0		0.1	2.0	**0.1**	**0.0**	2.7	0.9	0.6	12
Threonic acid-1,4-lactone	247	1,385	1.7		0.2	**2.8**	**0.1**	**0.0**	**0.6**	0.9	0.7	12
Dehydroascorbic acid	316	1,852	0.5		0.4	2.2	**0.1**	**0.1**	1.8	0.9	0.3	12
Aromatic Acids												
Benzoic acid[b]	179	1,253	1.9		1.7	1.6	2.4	2.8	1.9	–	0.5	1
Salicylic acid	267	1,514	3.1		1.0	**1.7**	**0.6**	**0.4**	0.9	3.1	1.1	7
p-Coumaric acid	293	1,948	4.7	5	–	–	–	–	–	**0.2**	**2.1**	8
N-Containing Compounds												
Putrescine	174	1,741	0.4		**3.3**	2.4	**2.9**	2.4	3.6	1.2	0.6	2
Urea	189	1,270	2.5		0.5	0.9	0.4	1.5	2.6	0.9	**8.3**	6
Allantoin	331	1,888	4.2	3, 4, 5	0.1	**2.6**	**0.0**	0.2	0.0	7.6	**30.7**	6
Uric acid	441	2,111	0.9		0.0	–	0.0	–	–	1.3	**69.4**	6
Sugars												
Rib	160	1,691	1.8		**2.3**	2.6	1.2	1.1	4.0	0.7	1.5	2
Raffinose	217	3,402	4.4		**1.8**	**1.7**	1.4	0.9	0.3	2.4	0.1	3
Xyl	160	1,670	1.3	3	0.0	**0.2**	**0.0**	0.7	**2.7**	0.8	**97.0**	6
Ara	160	1,676	2.1	3	0.1	**0.5**	**0.1**	0.7	1.0	0.6	**40.2**	6
Gal	160	1,892	0.5		0.1	**0.5**	**0.1**	0.5	**1.0**	0.4	**12.6**	6
Fru	307	1,885	0.6	3	**0.0**	**0.0**	**0.3**	**2.9**	**8.3**	0.9	**11.4**	6
Glc	160	1,916	0.8		0.2	**0.3**	0.5	**1.9**	**8.7**	1.0	**6.5**	6
Trehalose	191	2,751	2.2	5	0.2	0.4	**0.2**	0.7	2.5	3.3	**7.0**	6
Man	160	1,888	0.5		0.6	0.8	**0.6**	**0.8**	0.8	1.0	**2.6**	6
Fuc	117	1,747	0.4		1.3	1.4	1.0	1.7	1.3	1.0	2.6	6
Rha	160	1,728	0.3		0.5	**1.4**	**0.3**	**0.3**	1.3	**0.5**	**2.7**	8
Suc	451	2,653	1.0		0.6	0.6	1.0	1.3	**0.7**	1.1	**1.5**	9
Maltose	160	2,747	1.8		1.3	2.0	0.8	0.4	5.7	0.4	0.4	11
Polyols												
Sorbitol	319	1,937	1.2		2.1	1.8	1.8	1.6	2.9	0.8	0.8	2
Mannitol	319	1,929	1.2		**3.3**	2.9	2.0	1.3	7.1	0.5	0.7	2
Threitol	217	1,503	3.9		2.5	1.0	3.0	**2.9**	–	0.1	0.5	2
Glycerol	205	1,278	2.8		0.9	1.0	0.8	0.9	1.6	1.2	1.6	6
4-*O*-Methyl-*myo*-inositol, Ononitol	318	1,955	0.6		1.3	**5.0**	**0.3**	**0.2**	1.2	0.8	**3.6**	6
Galactitol	307	1,941	2.6		0.4	0.8	**0.4**	0.6	20.0	1.5	**3.9**	6
myo-Inositol	305	2,091	0.4		0.3	1.2	**0.2**	**0.1**	**0.9**	0.8	1.4	8
Erythritol	205	1,511	3.1		**0.1**	0.7	**0.1**	**0.2**	0.6	**1.8**	**3.6**	8
Galactinol	191	2,995	2.2		**0.4**	0.5	**0.5**	0.7	**0.5**	0.9	**2.2**	9
3-*O*-Methyl-D-*chiro*-inositol, D-Pinitol	231	1,835	2.2		0.4	0.8	**0.3**	**0.3**	**0.3**	0.7	**2.2**	9
Phosphates												
Glyceric acid-3-P	357	1,822	1.5		**7.1**	**19.8**	2.1	0.2	0.4	7.6	0.6	2
Man-6-P	160	2,324	2.2		1.1	**5.0**	**0.5**	**0.2**	0.6	**2.2**	**2.0**	5
Fru-6-P	315	2,324	2.4		1.1	**4.6**	0.6	0.2	0.4	5.4	1.5	5
Glc-6-P	387	2,337	1.7	5	1.5	**5.6**	0.6	**0.2**	0.4	**5.3**	1.8	5
Glycerol-3-P	299	1,777	3.8	4	**3.1**	**31.6**	1.3	0.1	0.1	108.0	1.8	5
myo-Inositol-P	318	2,430	1.2		0.6	1.0	**0.5**	**0.5**	**0.2**	**2.5**	1.5	7
Phosphoric acid	314	1,282	1.0		0.6	0.6	0.9	1.4	**0.3**	**4.9**	1.9	9
Unidentified												
–	216	1,763	1.9	4	2.4	1.3	**4.5**	5.6	**0.1**	3.2	0.5	1
–	71	1,601	3.5	3	1.8	2.2	1.3	1.2	1.4	0.8	1.6	2
[934; Pipecolic acid (2TMS)]	156	1,371	1.3	2	**14.5**	**52.4**	2.8	0.4	3.0	**2.9**	1.4	2

(Table continues on following page.)

Table I. *(Continued from previous page.)*

	m/z	RI, Median	RI, SD	Influence in PCA Component (Metabolite Lists among the Top 25 of Loadings Values)	Response Ratio (Nodule /Plant)	Response Ratio (Nodule /Root)	Response Ratio (Root, Nodule/ Shoot)	Response Ratio (Root/ Leaf)	Response Ratio (Lateral Root, Primary Root)	Response Ratio (Developing Leaf, Mature Leaf)	Response Ratio (Flower/ Other Organs)	Cluster Membership[c]
[910; Phenylpyruvic acid methoxamine (1TMS)]	250	1,602	2.0	2, 4	**40.3**	**16.4**	**46.2**	4.0	32.1	1.5	0.0	2
[824; 2-*O*-Glycerol-β-D-galactopyranoside (6TMS)]	263	2,190	3.1	5	**13.6**	**7.8**	**32.8**	7.2	1.0	3.6	0.0	2
[829; Melezitose (11TMS)]	361	3,389	3.0	2	**23.3**	**12.6**	**31.5**	5.2	11.0	1.2	0.0	2
[957; Suberylglycine (3TMS)]	188	1,638	4.4	1, 2, 3	**202.5**	**1138.6**	**129.2**	0.7	2.0	9.4	0.0	2
[802; Methylcitric acid (4TMS)]	243	1,930	2.0	2, 4	**59.2**	**42.0**	**34.0**	2.0	7.8	0.7	0.0	2
–	243	1,690	4.2	2	38.0	75.0	7.2	1.3	0.5	0.8	0.3	2
–	312	1,803	3.1	4	**2.8**	**16.3**	1.2	0.1	0.6	0.1	0.8	2
[630; DL-2-Methylcitric acid (4TMS)]	361	1,890	1.8	5	**2.8**	**2.2**	**2.3**	1.5	**0.1**	0.3	0.6	2
–	281	1,837	3.0		**2.0**	**2.0**	–	–	–	–	–	4
[877; Tetracosamethyl-cyclododecasiloxane][b]	279	2,758	4.4		0.7	0.2	10.3	32.3	–	0.2	0.2	4
[795; 3-Deoxy-arabino-hexaric acid (5TMS)]	245	2,115	1.6	1	**2.2**	**6.4**	0.8	**0.3**	**0.0**	0.9	**2.3**	5
[816; Hydroquinone-β-D-glucopyranoside (5TMS)]	254	2,607	0.9	5	**2.3**	**4.5**	0.7	**0.4**	1.1	**0.2**	1.8	5
–	142	1,624	3.6	5	1.1	**15.9**	0.4	1.6	2.7	2.8	**7.3**	6
[632; Pro (2TMS)]	186	1,594	2.6	1	0.0	0.8	**0.0**	**0.1**	0.8	3.9	**49.7**	6
[910; 4-*O*-D-Glc-β-D-glucopyranoside (8TMS)]	169	3,068	7.1	5	–	–	0.1	0.5	0.3	0.2	**21.3**	6
[607; L-Asp (3TMS)]	232	1,957	1.8	1, 4	0.3	1.0	**0.2**	0.3	**0.0**	10.4	**6.7**	6
[799; Maltose (8TMS)]	361	2,226	2.6	2	**0.0**	–	**0.0**	–	–	2.0	**6.4**	6
[846; 1-Methyl-β-D-galactopyranoside (4TMS)]	205	2,102	4.9	1, 4	0.3	0.7	**0.3**	0.5	**0.0**	1.8	**4.2**	6
[674; Gln (4TMS)]	301	1,597	3.8	4	0.9	16.4	0.5	–	0.0	–	2.2	6
[817; Glc-6-P methoxy-amine (6TMS)]	299	2,569	3.7	5	0.3	2.3	**0.2**	0.1	0.4	3.8	0.9	7
[787; Trehalose (8TMS)]	361	2,597	3.1	1, 5	0.1	**5.0**	**0.0**	**0.0**	1.0	**3.1**	1.1	7
[866; Gulose (5TMS)]	364	2,169	2.3	4	0.1	0.4	**0.1**	**0.1**	**0.2**	2.6	1.1	7
[746; Gulose (5TMS)]	204	2,431	4.4	4	-	-	**0.1**	**0.1**	0.2	**0.7**	1.6	8
[810; L-Rha (4TMS)]	249	2,188	2.4	2	**0.0**	1.6	**0.0**	0.0	0.0	**0.2**	**2.6**	8
[802; Gulose (5TMS)]	159	2,443	2.5	1	**0.1**	**2.2**	**0.0**	**0.0**	1.7	0.4	**4.2**	8
[814; Ribonic acid (5TMS)]	333	1,762	0.4		**0.1**	**0.5**	**0.2**	**0.2**	1.6	0.7	**3.1**	8
[914; Ribonic acid (5TMS)]	333	1,774	0.5		**0.2**	**0.6**	**0.2**	**0.2**	1.3	0.7	**2.6**	8
[849; 1-Methyl-β-D-galactopyranoside (4TMS)]	174	2,161	3.1	1	0.1	0.9	**0.1**	**0.1**	**0.1**	1.6	2.3	8
[797; Gulose (5TMS)]	91	2,411	2.4	1	0.1	**2.4**	**0.0**	**0.0**	**0.8**	1.5	**3.7**	8
[841; 1-Methyl-β-D-galactopyranoside (4TMS)]	230	2,169	2.5	1	0.1	0.3	**0.2**	**0.3**	**0.0**	1.3	**2.5**	8
[649; L-Ala (2TMS)]	132	1,408	2.3	2, 4	**0.0**	0.0	**0.4**	**0.6**	**0.3**	**1.9**	**2.5**	9
[953; Malonic acid (2TMS)]	233	1,213	4.4	1, 3, 4, 5	0.0	**17.2**	0.0	0.0	5.9	0.3	0.4	11

(Table continues on following page.)

Table I. *(Continued from previous page.)*

	m/z	RI, Median	RI, SD	Influence in PCA Component (Metabolite Lists among the Top 25 of Loadings Values)	Response Ratio (Nodule /Plant)	Response Ratio (Nodule /Root)	Response Ratio (Root, Nodule/ Shoot)	Response Ratio (Root/ Leaf)	Response Ratio (Lateral Root, Primary Root)	Response Ratio (Developing Leaf, Mature Leaf)	Response Ratio (Flower/ Other Organs)	Cluster Membership[c]
[624; Xylulose (4TMS)]	306	1,590	1.2	5	0.2	**2.6**	**0.1**	**0.0**	**2.4**	**0.1**	0.9	11
[827; Gulose (5TMS)]	204	2,678	5.5	4	-	-	0.0	0.0	1.3	0.0	0.2	11
[840; Melibiose (8TMS)]	217	2,456	1.9	1	0.0	0.7	**0.0**	**0.0**	**0.4**	**0.1**	1.6	11
[791; 4-*O*-D-Glc-*β*-D-galactopyranoside (8TMS)]	247	2,510	5.3	5	0.1	0.7	0.1	0.1	2.2	**0.1**	0.4	11
[716; 4-*O*-D-Glc-*β*-D-galactopyranoside (8TMS)]	361	2,950	4.0	4	0.0	0.7	**0.0**	**0.0**	0.0	0.0	0.6	11
–	204	2,189	2.0	1, 4	0.6	**94.9**	**0.1**	**0.0**	0.2	**0.3**	1.6	11
[851; Gulose (5TMS)]	235	2,191	1.2	2	0.0	1.4	**0.0**	0.0	1.3	**0.2**	1.0	11
–	117	2,611	2.2	1	0.2	2.0	**0.1**	**0.1**	0.0	**0.2**	1.2	11
[690; Raffinose (11TMS)]	361	2,525	2.5	5	0.8	**4.2**	0.6	0.1	-	**0.1**	0.2	11
[752; 1-Methyl-6-deoxy-galactoopyranoside (3TMS)]	204	2,071	2.3	5	2.0	**4.7**	0.7	0.3	**0.2**	**0.1**	0.5	11
[625; 2,2,7,7-Tetra-methyl-4,5-diphenyl-3,6-dioxa-2, 7-disilaoctane][a]	179	2,261	2.7	1, 3, 5	**0.0**	9.9	**0.0**	**0.0**	0.8	1.2	0.7	12
[919; Arabino-Hexos-2-ulose-bis(dimethyl-acetal) (4TMS)]	234	1,485	2.1	3, 5	0.0	0.8	**0.0**	**0.0**	0.6	1.5	0.1	12
[827; Suc (8TMS)]	450	2,713	3.5	2	0.0	0.9	**0.0**	0.0	-	0.5	0.3	12
[849; 4-*O*-D-Gluco pyranose-*β*-D-galactopyranoside (8TMS)]	204	2,726	4.7	3	0.1	0.5	**0.1**	**0.1**	0.1	0.6	0.3	12

[a]Combined quantitative information due to chemical interconversion. [b]Noncorrected artifacts for the detection of jet unknown artifact compounds. [c]For cluster description, refer to Figures 6 and 7 (cluster numbers equal metabolite classes of Figs. 6 and 7).

address fragments that belong to unidentified MSTs, we use the following nomenclature: match value, and substance name of best fit, separated by a semicolon and set into brackets, e.g. 243_1930_[802; Methylcitric acid (4TMS)] (e.g. Fig. 4).

MST-Based Identification of Metabolites in Lotus

Comparison of MSTs derived from Lotus organs with those of pure reference compounds enabled the identification of 87 compounds among the hundreds represented on GC-MS chromatograms (Table I). These included most of the common amino acids as well as polyamines; many organic acids, including TCA cycle intermediates; aromatic acids; sugars and sugar phosphates; and polyols (Table I). A number of likely chemical contaminants, from human or laboratory sources, or reagent impurities, including lactic acid, benzoic acid, and oligomethyl-cyclosiloxanes, were also identified.

A small set of MSTs was found to represent more than one metabolite. For example, pyroglutamic acid is formed from Gln, and, to a lesser extent, from Glu during extraction and derivatization of metabolites. However, the classification of Gln, Glu, and pyroglutamic acid into different clusters (see Fig. 6) indicated minimal cross-contamination in this analysis. Arg and citrulline may be converted completely into Orn during chemical derivatization. In our analyses, no specific derivatives of Arg or citrulline were found. Thus, the MST of Orn represented the sum of endogenous Arg, citrulline, and Orn.

Numerical and PCA Analysis of Organ Metabolic Phenotypes

Manual inspection of GC-MS chromatograms indicated major similarities in metabolism of developing and mature leaves, as well as similarities between lateral and primary roots (Fig. 1). To analyze similar-

ities and differences numerically, we performed automated peak integration using 1,046 mass spectral fragments, representative of about 500 MSTs. MSTs representing known or unknown metabolites were analyzed, as a rule, using one to four specific mass spectral fragments within the respective retention time window (see "Materials and Methods" section "Generation of a Metabolite Response Matrix"). Choice of fragment mass and retention time window was performed manually and was facilitated by nonsupervised collection of MSTs (Colebatch et al., 2004). Thus, a large matrix of 1,046 fragment responses, which describe 64 samples from 6 organs of *L. japonicus*, was generated. PCA (Joliffe, 1986) was applied to gain insight into the nature of the above multivariate data. PCA identifies and ranks major sources of variance within data sets and allows clustering of biological samples into both expected and unexpected groups based on similarities and differences in the measured parameters. PCA also identifies those data elements, e.g. MSTs representing known or unknown metabolites, which contribute most to each of the principal components that describe the variance in metabolite profiling data sets (Roessner et al., 2001a, 2001b). Finally, if sample classes can be clearly distinguished when projected onto any of the principal components, PCA enables identification of those MSTs and metabolites that distinguish the sample classes.

The first 5 principal components derived from the above data matrix encompassed 77.3% of the total variance from this data set (Fig. 2). The first component accounted for 37.1% of the variance and allowed distinction of shoot organs from root organs (Fig. 2A). Nodules exhibited more similarity to roots than to shoot organs according to the first component. However, the second component, which comprised 17.8% of the variance, demonstrated that the data set contained metabolite measurements that distinguished between nodule and root profiles (Fig. 2A). Subsequent principal components revealed other differences between the various organs. Thus, the third and fourth components, encompassing 11.2% and 7.3% of the variance, respectively, indicated that general markers for flowers and primary roots exist (Fig. 2B). The fifth component clearly separated developing leaves from other organs (Fig. 2C). No subsequent components allowed a clear discrimination between sample types (e.g. Fig. 2C, component 6).

The organ samples described above were harvested at one developmental stage, but at different times of a single day/night cycle. No principal component was found that reflected diurnal changes in metabolism. This finding indicated that diurnal changes resulted in only minor changes in metabolite profiles compared to those resulting from organ development and differentiation. PCA analysis of leaf samples only indicated some diurnal changes in this organ (data not shown). However, the small number of samples from each time point prohibited identification of significant shifts in leaf metabolism during the diurnal cycle.

To test the robustness of PCA in distinguishing between different organs, we analyzed GC-MS data from plants harvested over a 10-week period (7–17 weeks after germination), following growth under different culture conditions in different seasons (Fig. 3A). This data set was expected to be more variable than the first. In fact, this appeared to be the case, and PCA analysis proved less successful in distinguishing between samples of different organs (Fig. 3A). Nonetheless, nodule and root samples were mostly separated from shoot organs by component 1 of PCA, which accounted for 22.9% of the total variance. The 2 subsequent components covered a sum of 14.6% of variance but did not yield distinctions between the samples that could be linked to organ age or plant growth conditions. However, the fourth component, which encompassed 5.5% of the variance, separated nodule samples from those of other organs (Fig. 3A).

PCA Analysis Reveals Specific Metabolites That Distinguish Different Organs

The contribution of each metabolite to a specific component is reflected by the loading value derived from PCA analysis. Those metabolites with highest loading values are indicated to have the strongest influence on the respective characteristics of a component. We focused on the loading values of components 1 and 2 of experiment 1 (Fig. 2). The 25 most influential fragment masses of each component were analyzed (Fig. 4, A and B). The first component, which described the root to shoot difference, was influenced most by Pro, Gln, erythronic acid, and glucaric acid. The second component, which revealed differences between nodules and other organs, was influenced most by Asn, Gln, Glu, Trp, and octadecanoic acid. Multiple MSTs of unidentified compounds were also found to contribute substantially to components 1 and 2. We selected MST [802; Methylcitric acid (4TMS)] and an identified metabolite, glucaric acid, to demonstrate the possibility of gaining biological insight about a compound, even if its chemical identity is unknown. The choice of these two compounds was made with reference to data from the second experiment described above (Fig. 3). Glucaric acid was found to be important for root and shoot distinction: fragment masses 292, 333, and 373 at RI 2,014 (Fig. 3B). The unknown MST [802; Methylcitric acid (4TMS)] was also found to be a reproducible marker of nodules: fragment 243 at RI 1,929 (Fig. 3B).

Further Analysis of an Unidentified Metabolite

MST [802; Methylcitric acid (4TMS)], as the nomenclature indicates, was found to be highly similar to a typical bacterial metabolite, 2-methylcitric acid. This match was found in the commercially available NIST02 mass spectral library (Ausloos et al., 1999). In contrast, glucaric acid was immediately identified by mass spectral match with a custom set of replicate

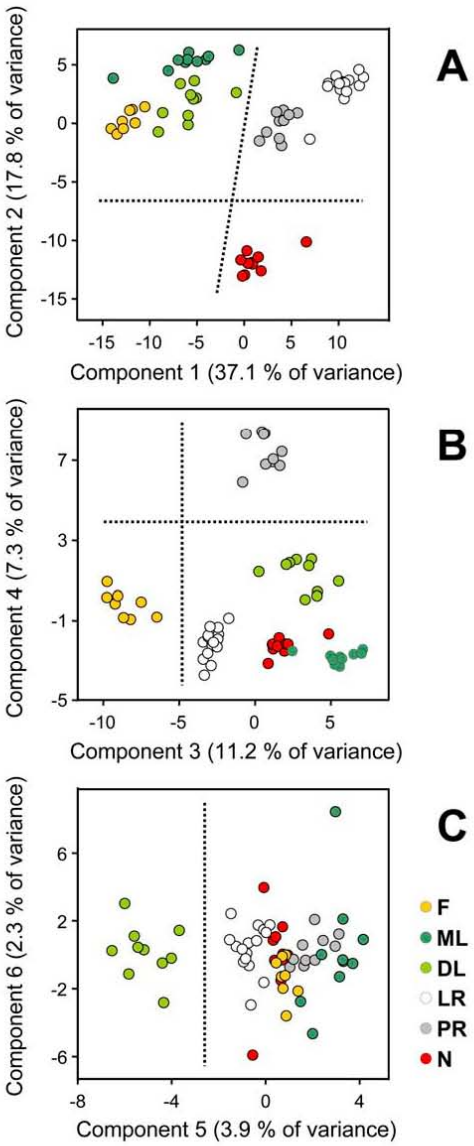


Figure 2. PCA analysis of GC-MS profiles, which represent polar metabolites of *L. japonicus* organs harvested in the course of 1 d at 12 weeks after germination. Samples were projected into three bi-plots of principal components that were arranged in descending order of variance. Each of the first five principal components allowed clear distinction of metabolite profiles from specific organs. Component 1 separated root from shoot organs, component 2 described the difference of nodules as compared to plant organs, and components 3 to 5 described the distinction of flowers, primary roots, and developing leaves from the remainder of the samples. LR, Lateral roots (n − 15); PR, primary roots (n − 10); F, flowers (n − 8); DL, developing leaves (n − 10); ML, mature leaves (n − 10); and N, nodules (n − 10).

mass spectra (match value = 944 to 989) and by low RI deviation (ΔRI = 0.3 to 0.5), as described in (Wagner et al., 2003). Typical mass spectra of both compounds contained fragment masses that were used for quantification (Fig. 5A). In an attempt to identify the unknown MST, standard addition experiments were performed with commercially available DL-2-methylcitric acid. This compound generated an MST with mass spectral similarity (match value = 779) but high RI deviation (ΔRI = 89.0). Thus, we confirmed by similarity that MST [802; Methylcitric acid (4TMS)] belongs to the class of methylcitric acids, but we cannot currently define the specific structural position of the methyl group. In addition, we were able to identify true DL-2-methylcitric acid by RI and mass spectral match (700 to 842 and ΔRI = 0.2 to 0.6) with a hitherto unknown MST from the custom MS and RI library of *L. japonicus* (Colebatch et al., 2004).

Despite the lack of a specific structure for MST [802; Methylcitric acid (4TMS)], we investigated the distribution pattern of the underlying metabolite in Lotus organs (Fig. 5B). MST [802; Methylcitric acid (4TMS)] was found at high levels in nodule samples, while all other organs contained only traces of this compound. In contrast, DL-2-methylcitric acid was relatively high in nodules and flowers but low in lateral roots (Fig. 5B). These results indicate MST [802; Methylcitric acid (4TMS)] is a good marker substance for nodules, while DL-2-methylcitric acid is more evenly distributed throughout the plant. Furthermore, we found glucaric acid to be low in roots and nodules but high in leaves and flowers. Thus, this compound was confirmed as a good marker for shoot organs.

Analysis of Metabolite Distribution Patterns

As illustrated in Figure 5B, metabolites were found to exhibit specific distribution patterns in the organs of Lotus. We applied HCA to the MST distribution of all 87 identified and 49 unidentified compounds that were among the top 25 most discriminatory from PCA analysis (Figs. 1 and 4). Only one mass fragment was used for each MST in this analysis. Following HCA, we grouped the MSTs into 12 classes (Fig. 6). Two of the classes, 1 and 4, were occupied by known laboratory contaminants and were excluded from further analysis. The properties of the remaining 10 classes were investigated further (Fig. 7).

Class 2 contained metabolites that were present at relatively high levels in nodules and at low or intermediate levels in other organs. MST [802; Methylcitric acid (4TMS)], like Glu, Asn, putrescine, and mannitol, were characteristic members of this class. Class 3 compounds, with relatively high levels in

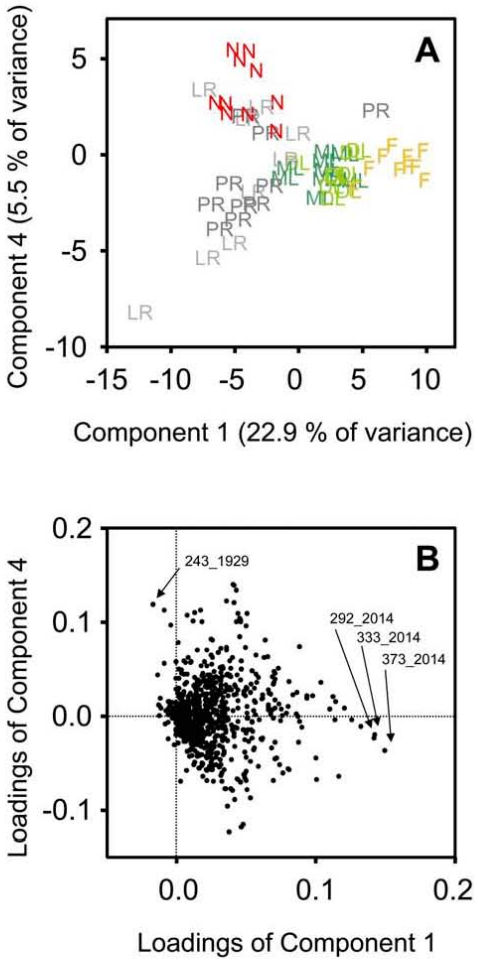


Figure 3. PCA analysis of a second set of GC-MS profiles, which represent polar metabolites of *L. japonicus* organs that were harvested in the course of a 6-month period at random stages 7 to 17 weeks after germination. PCA analysis of this data set confirmed dominating influence of the root-to-shoot differentiation on the variance observed in GC-MS profiles. In addition, nodule-to-plant differences were detectable in the sample scores plot (A). Loadings analysis (B) confirmed importance of glucaric acid, represented by fragment masses 292, 333, 373 at RI − 2,014 for component 1 and importance of MST [802; Methylcitric acid (4TMS)], represented by fragment mass 243 at RI − 2,014, for component 4. Use of fragments may change from data set to data set, because changes in metabolite levels cause fragments of low abundance to drop below detection limits.

nodules and leaves, had only 2 members, raffinose and 2-ketoglutaric acid. Class 5, which had high levels in nodules and flowers, comprised DL-2-methylcitric acid and 10 other compounds, including Glc-6-P and 2-aminoadipic acid. Class 6 was the dominant metabolite class and comprised metabolites with high levels in flowers only, for example Pro, Val, Trp, ononitol, and Gln. Classes 7 and 8 were similar and contained metabolites enriched in shoot organs, such as glucaric acid. However, class 8 contained metabolites that had high levels in all above-ground organs, while class 7 metabolites had reduced levels in mature leaves. Class 9 metabolites exhibited relatively high levels in shoot organs and in primary roots. Class 10 contained metabolites with low levels in nodules only. Class 11 and 12 metabolites exhibited high levels in mature leaves. While metabolites of class 12 were also present at high levels in developing leaves, class 11 metabolites had only low or medium levels in other organs. Detailed information on the class membership of each MST, together with short descriptions of each class, is included in Table I.

ANOVA

PCA analysis pointed to metabolites that may be important for organ differentiation. Metabolite clustering by HCA resulted in a rough overview of general metabolite distribution patterns. As an extension to these analyses, ANOVA was used to assess the statistical significance of differences in the distribution of each metabolite. Seven comparisons of organs and groups of organs were performed. These comparisons were motivated by sample classifications made evident by PCA analysis: (1) comparison of nodule with all other plant samples; (2) comparison of nodule with root samples; (3) comparison of below-ground with above-ground samples; (4) comparison of root samples with shoot, including flower samples; (5) comparison of flower samples with all other samples; (6) comparison of lateral and primary roots; and (7) comparison of developing and mature leaves. Differences in metabolite levels were calculated as ratios and compiled in Table I. Differences in metabolite levels that were significant at $P \leq 0.01$ are indicated in the table.

Amino acids exhibited two major sites of accumulation: nodules and flowers. Asn, homoSer, Glu, and Cys levels were significantly higher in nodules than in other organs (Table I). In contrast, most other amino acids, especially Trp, Pro, Leu, Val and Gln, were enriched in flowers. Most amino acids were present at higher levels in leaves than in roots. Developing leaves contained higher concentrations of most amino acids than mature leaves, although most differences were not significant.

Only 4 identified organic acids accumulated significantly in nodules compared to all other organs: octadecanoic acid, threonic acid, gluconic acid, and 2-methylcitric acid. Like amino acids, most organic acids were present at significantly lower levels in roots than in leaves. Some organic acids, for example GlcUA and quinic acid, were highly enriched in flowers. Massive accumulation of the N-containing compounds, uric acid, allantoin, and urea was also found in flowers.

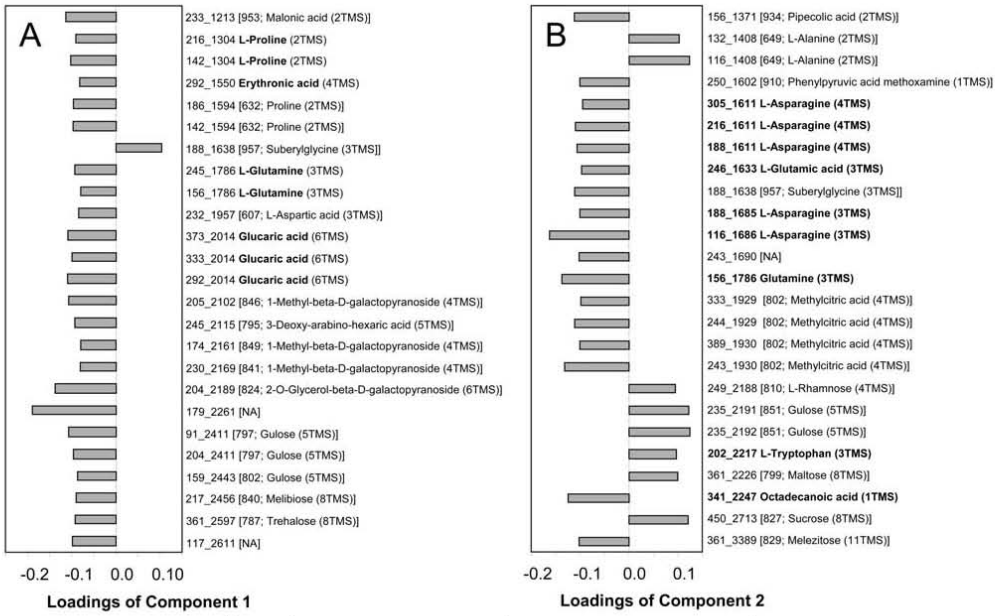


Figure 4. The 25 most influential fragment masses, which contribute to component 1 (A) and 2 (B) of Figure 2A. The top ranking fragment masses were sorted according to RI, and MSTs as well as metabolite names were manually assigned. MSTs representing metabolites and unidentified MSTs were represented by more than one fragment mass for the purpose of validating metabolite influence. MSTs representing identified metabolites were labeled by fragment mass, RI, and name of metabolite derivative. For example, Asn was found to influence component 2 with 2 MSTs, resulting from different degree of chemical derivatization, namely L-Asn (4TMS) and L-Asn (3TMS). These MSTs were measured by fragment masses 188, 216, 305 at RI − 1,611, and masses 116, 188 at RI − 1,686. Unidentified MSTs (brackets) are characterized by mass spectral match value, name of best match, and degree of silylation (parentheses).

In contrast, putrescine levels were higher in nodules and roots than in other organs (Table I).

Sugars exhibited variable organ distribution patterns. Most striking was the accumulation of Xyl, Ara, and Gal in flowers. Nodules exhibited significant accumulation of Rib and raffinose compared to other organs. Whereas most sugars were present at similar levels in roots and leaves, Glc and Fru were significantly higher in roots and especially in lateral roots. Sugar alcohols and sugar phosphates were distributed more evenly throughout the plant but exhibited a tendency to be low in roots. Only mannitol, glyceric acid-3-P, and glycerol-3-P exhibited significant accumulation in nodules. Ononitol, pinitol, and galactitol were relatively high in flowers. Developing leaves accumulated a range of phosphorylated compounds, especially phosphoric acid and Glc-6-P.

Most of the MSTs of unidentified compounds, which were selected from the top-ranking loading values of PCA components 1 to 5, exhibited significant differences that substantiated their importance as markers for nodules, roots, or shoot organs. Among these were MSTs that showed the most extreme changes between organs, such as MSTs [957; Suberylglycine (3TMS)] and [802; Methylcitric acid (4TMS)] (Figs. 5 and 7).

DISCUSSION

Current methods of metabolome analysis are far from comprehensive. We selected the GC-MS based method reported earlier (Fiehn et al., 2000; Roessner et al., 2000), which allows analysis of the low to medium M_r, soluble, polar metabolic complement and comprises primary metabolites, such as 24 major and minor amino acids; 29 hydroxylated, nonhydroxylated and aromatic organic acids; 4 amines and amides; 13 mono-, di-, and tri-saccharides; 10 polyols; and 7 phosphorylated compounds (Table I). The choice of this specific GC-MS approach was motivated by the almost complete coverage of those metabolite classes that have received attention in past studies of SNF, namely amino acids, carbohydrates, and organic acids. This enabled validation of some of the GC-MS data by comparison with published data.

Although not as comprehensive as transcript profiles derived from whole-genome oligonucleotide ar-

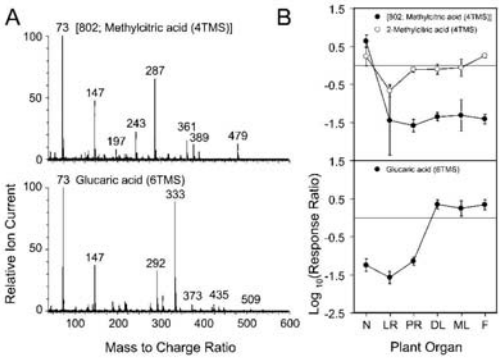


Figure 5. Mass spectra (A) and distribution (B) of MST [802; Methylcitric acid (4TMS)] (black circles), DL-2-methylcitric acid (white circles), and glucaric acid (black circles at bottom). Metabolites were identified by mass spectral match and chromatographic retention. Glucaric acid had match values 944 to 989 (ΔRI = 0.3 to 0.5). MST [802; Methylcitric acid (4TMS)] was similar to DL-2-methylcitric acid but had high RI deviation, match value 779 (ΔRI = 85). The match of DL-2-methyl citric acid was 700 to 842 (ΔRI = 0.2 to 0.6).

rays, metabolite profiles allow similar insights at the metabolic level. Biological samples can be classified according to their metabolic phenotype, e.g. the quantitative and qualitative make-up of the metabolome (Figs. 2 and 3); and metabolites, like gene transcripts, can be classified according to their distribution within the various samples under investigation (Figs. 6 and 7). Thus, sets of metabolites can be identified that are not only of diagnostic value but also may indicate the function of metabolites in certain organs or under certain conditions. Below, we discuss three aspects of metabolite profiling: (1) analytical and technical aspects; (2) the data mining strategy applied in this study; and (3) the potential for new insight into a biological process such as SNF that is afforded by comparative metabolomics.

Analytical and Technical Issues

The metabolome of an organism is subject to rapid, enzyme-catalyzed change in response to environmental as well as endogenous factors. Photosynthetic organisms like plants, for instance, undergo profound diurnal changes in metabolism that are related to the light-dark cycle. It is important to take this into account by standardizing growth conditions and synchronizing harvesting times. Equally important is the fact that enzymatic and nonenzymatic conversion of a metabolite to one or more products does not necessarily cease at the time of harvest. Precautions must be taken to avoid such conversions during harvesting, storage, extraction, and derivatization of metabolites (Kopka et al., 2004). For instance, we rapidly harvested plant material in liquid nitrogen, stored at −80 C, and extracted in a mixture of solvents that avoids as much as possible metabolite degradation or modification. To judge the effects of uncontrolled growth conditions and harvest times on metabolite profiles, we compared our well-controlled experiment (Fig. 2) with a less controlled one (Fig. 3). Variance due to experimental intervention is generally much greater than the purely analytical variation, which in the case of metabolite profiling typically ranges from 10% to 25% relative SD, but can be as low as 1% to 5% (Fiehn et al., 2000; Roessner-Tunali et al., 2003).

Data Mining and Analysis

Metabolite profiling data from GC-MS are highly complex, which presents challenges for identification and quantification of metabolites. Unfortunately, bioinformatics tools for automated processing of data are less well-developed for metabolomics than for proteomics and transcriptomics. Currently, the full nonredundant inventory of metabolites from one set of GC-MS profiles cannot be assessed without time consuming manual curation. For this reason, we produced so-called nonsupervised collections of redundant mass spectra from automatically generated mass spectral deconvolution of representative chromatograms (Wagner et al., 2003) for cross-referencing of results and MST identity from *Lotus* (Colebatch et al., 2004). Known metabolites can be identified by comparing RI and mass spectra of pure standard compounds with mass spectra found in preparations of plant samples. Unidentified MSTs can be retrieved from these libraries for further analysis if there are indications that they may be important. We briefly discuss our current approach to extract useful information from metabolic data matrices below.

A preliminary overview of general similarities and differences between samples is a useful first step in data analysis. Visual inspection of chromatograms is insufficient for this purpose (Fig. 1A). HCA and PCA have been widely applied for data reduction to avoid the need to check each single metabolite for relevant changes (e.g. Roessner et al., 2001a, 2001b). HCA sorts and classifies according to the degree of similarity between metabolite profiles. HCA analysis of our profiles confirmed that each organ of *L. japonicus* had a characteristic metabolic phenotype (Fig. 1B). PCA of the same data set confirmed HCA results but proved to be superior to HCA in allowing determination not only of differences between sample classes, but also of ranking these differences according to the portion of comprised variance (Fig. 2). Thus, we were able to demonstrate that root-to-shoot differences were responsible for a major part of the variance in the combined data set, followed by differences between nodules and other organs, etc. In the same way, we were able to conclude that intraorgan diurnal changes were relatively small compared to interorgan differences. Nonetheless, diurnal changes were evident in a focused analysis of leaf samples only, which was not

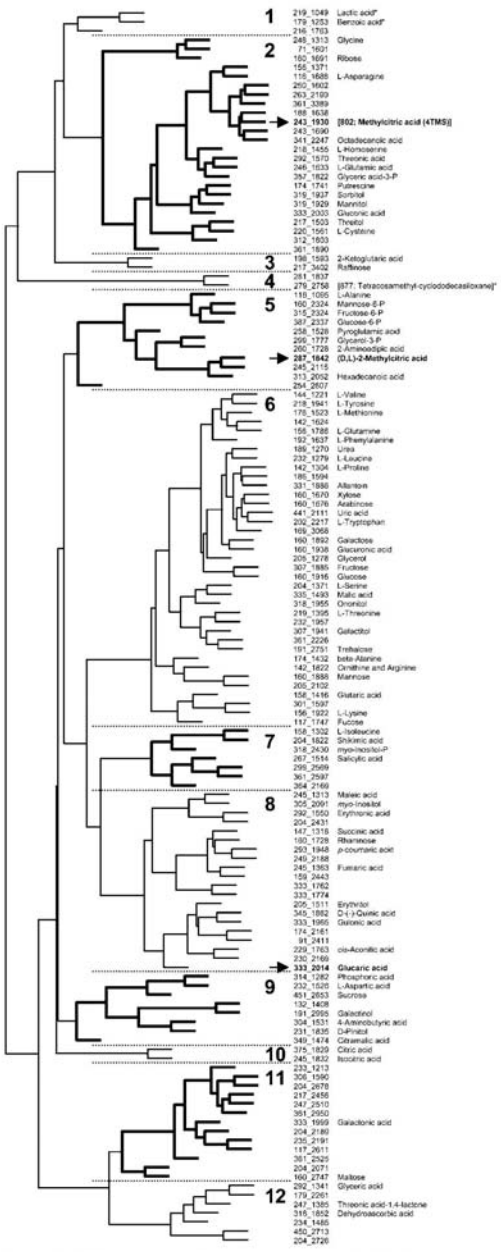


Figure 6. HCA analysis of metabolite and MST distribution in *L. japonicus* organs. HCA sorts and classifies metabolites according to their distribution pattern. Twelve classes were identified manually (broken lines). Classes 1 and 4 comprised typical laboratory contaminations and allowed to exclude one MST, fragment 281 at RI = 1,837, from further analysis. Class membership and descriptions may be found in Table I. Typical examples of metabolite distributions are presented in Figure 7. Cluster descriptions are as follows.

2: N (high), F (low-high), others (low)
3: N, DL, ML (high), others (low-high)
5: N, F (high), others (low-high)
6: F (high), others (low-medium)
7: Root (low), DL, F (high), ML (medium)
8: Root (low), Shoot (high)
9: N (low), LR (low), others (medium-high)
10: N (low), others (medium-high)
11: ML (high), others (low-medium)
12: ML (high), DL (medium-high), F, Root (low-medium)
1: Test for solvent contamination
4: test for reagent artifact.

presented here because the low number of replications per time point prohibited statistical rigor. Another advantage of PCA over HCA is the possibility to derive a list of metabolites that contribute to each principal component. If principal components separate different sample groups, e.g. components 1 and 2 that separate nodules from all other organs (Fig. 2), a rank-ordered list of MSTs representing known and unknown compounds that distinguish between the groups can be obtained (Fig. 4). In this way, we identified 136 MSTs from among the initial 500 MSTs, representing 87 known and 49 unknown compounds, which distinguished between sample groups. Obviously, such a list also provides insight into the metabolic differences between distinct sample groups such as organs. Thus, by identifying the most striking features of a profiling data set, PCA is an efficient first step of the data mining process.

From the set of 136 MSTs that were identified by PCA to have potential diagnostic properties, we selected 2 MSTs representing a known (glucaric acid) and an unknown compound [802; Methylcitric acid (4TMS)] for further analysis (Fig. 5A). Validation of the diagnostic value of these metabolites was performed via PCA analysis of a second experiment, in which plants were grown under more variable conditions. Once again, both compounds were among the most influential metabolites separating root from shoot, and root from nodule samples, respectively (Fig. 3B). A subsequent analysis of the distribution pattern (Fig. 5) clearly supported the diagnostic properties of the selected compounds. After having established compound relevance we manually confirmed compound identity for glucaric acid and found mass spectral similarity of MST [802; Methylcitric acid (4TMS)] to methylcitric acid. Subsequent standard addition experiments with commercially available DL-2-methylcitric acid confirmed this similarity. As a by-product of this work, we discovered true DL-2-methylcitric acid among the MSTs of unidentified compounds and were able to add this novel identification and its distribution pattern to the set of fully characterized MSTs (Fig. 5B). In the absence of additional commercially available candidate reference substances, further attempts to

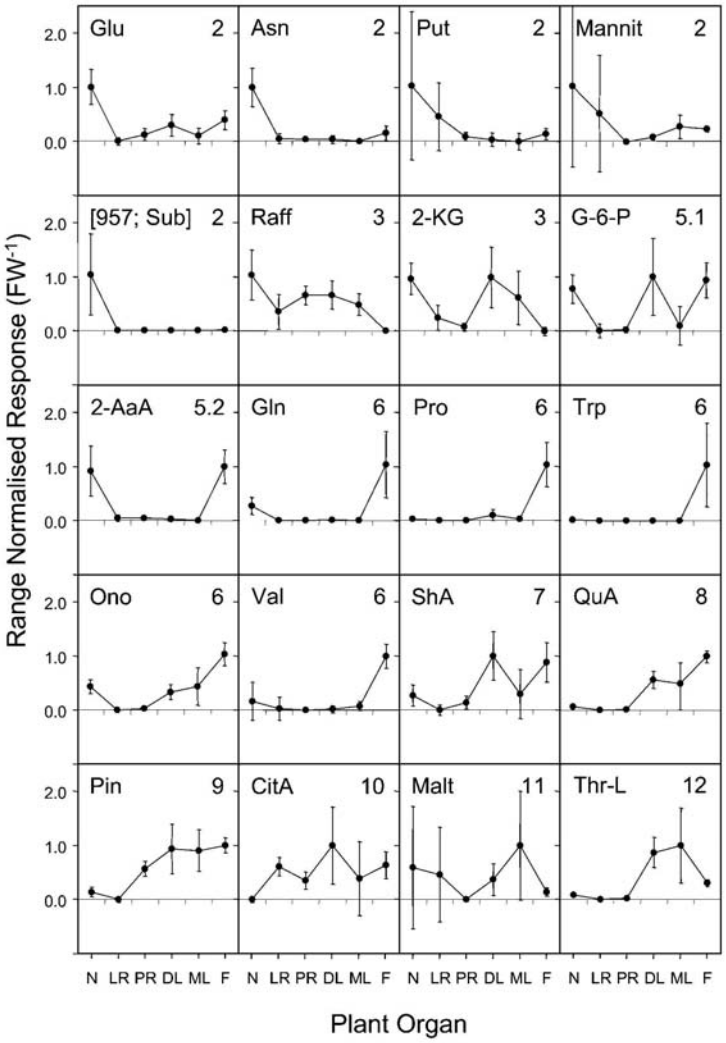


Figure 7. Typical distribution patterns of representative metabolites from HCA classes. Metabolites were range normalized to minimum 0 and maximum 1 for comparative purpose; error bars represent SDs. Class membership is indicated in top right corner. 246_1633_L-Glu, 116_1686_L-Asn, 156_1786_L-Gln, 142_11304_L-Pro, 174_1741_putrescine (Put), 319_1929_mannitol (Mannit), 318_1955_ononitol (Ono), 144_1221_L-Val, 188_1638_[957; SuberylGly (3TMS)] ([957; Sub]), 387_2337_Glc-6-P, 260_1728_2-aminoadipic acid (2-AaA), 202_2217_L-Trp, 217_3402_raffinose (Raff), 198_1593_2-ketoglutaric acid (2-KG), 204_1822_shikimic acid (ShA), 345_1862_D-(-)-quinic acid (QuA), 231_1835_D-pinitol (Pin), 375_1829_citric acid (CitA), 160_2747_maltose (Malt), and 247_1385_threonic acid-1,4-lactone (Thr-L). The nomenclature of fragment mass, RI, and MST names is defined in paragraph "Analysis and Nomenclature of MST."

identify MST [802; Methylcitric acid (4TMS)] will require time-consuming chemical purification of the compound and structural characterization, by NMR analysis, for example.

HCA analysis was applied to all 87 MSTs representing known metabolites and to the 49 representing unidentified compounds, which were found to be interesting from PCA analysis (Fig. 6). HCA demonstrated that the distribution patterns of glucaric acid and MST, [802; Methylcitric acid (4TMS)] were not unique but that both metabolites belonged to groups of metabolites with similar distribution, namely class 2 comprising 24 nodule-enriched metabolites, which were enriched in this organ by between 2- and 200-fold, and class 8 comprising 20 shoot enriched metabolites, which were depleted in roots and nodules by between 2- and 50-fold (Table I). Moreover, HCA supported distribution patterns that were indicated by PCA; e.g. 13 flower-enriched metabolites of class 6 accumulated by more than 10-fold (Table I; Fig. 7). Although HCA allowed rough classification, ANOVA was required to distinguish between those metabolites that were significantly enriched and those that were not. Table I comprises all relevant comparisons that

were performed and demonstrates a multitude of significant metabolite enrichments with factors >10-fold or <0.1-fold.

Nodule Metabolism: Some Insights from GC-MS Analysis

PCA analysis revealed that many compounds were enriched in nodules compared to other plant organs, including Asn, Glu, Gln, homoSer, Cys, putrescine, mannitol, threonic acid, gluconic acid, glyceric acid-3-P, glycerol-3-P, and octadecanoic acid. Some of these differences were expected and confirm what is known about nodule metabolism. For instance, SNF is a source of ammonium for amino acid biosynthesis and many legumes, including Lotus export fixed nitrogen in the form of amines, especially Asn and Gln (Vance et al., 1987). Therefore, it was reassuring to find these amino acids at higher levels in nodules than in roots or in the plant as a whole (Table I). In a similar vein, it is known that glycolysis is enhanced in nodules compared to roots (Copeland et al., 1989; Day and Copeland, 1991), and this was reflected by the ratio of hexoses to hexose-phosphates in these organs. The relative abundance of both Fru-6-P and Glc-6-P were about 5 times higher in nodules than in roots, while Fru and Glc were much lower in nodules than in roots. These changes are indicative of increased flux through glycolysis (Roessner et al., 2001; Fernie et al., 2002), even though metabolite levels per se are not a direct measure of flux.

A number of compatible solutes, which typically accumulate in plants in response to osmotic stress, were found to be elevated in nodules compared to roots and other organs. These included the polyols, ononitol, mannitol, and sorbitol; the amino acid, Pro; and the polyamine, putrescine (Table I). Accumulation of osmoprotectants in nodules may indicate that cells in this organ are subject to osmotic stress. Hypoxia, which can cause osmotic stress in plant cells via effects on water uptake and loss (Nuccio et al., 1999), could be responsible for this build-up of compatible solutes. Interestingly, genes encoding putative mannitol transporters are among those induced during nodule development (Fedorova et al., 2002; Colebatch et al., 2004), and these may be involved in importing polyols derived from photosynthesis in the shoot (Noiraud et al., 2001). On the other hand, a proteomic study of the Lotus identified a putative mannitol transporter on isolated peribacteroid membrane/SM (Wienkoop and Saalbach, 2003), which indicates that polyols may be transported between the plant and bacteroids. Sorbitol dehydrogenase, which interconverts D-Fru and D-sorbitol, is induced in nodules (Colebatch et al., 2004), indicating that de novo synthesis of polyols may also occur in this organ. Genes involved in Pro and polyamine biosynthesis are also induced during nodule development, which could account for accumulation of these compounds (Colebatch et al., 2004; Flemetakis et al., 2004). While this data may indicate that osmotic stress is a normal aspect of nodule physiology, a more trivial explanation would be that our sand-grown plants were generally water-stressed at the time of harvest. However, this explanation is at odds with the observation that roots contained significantly lower levels of specific compatible solutes than did nodules of the same plants.

Relatively high levels of Cys were found in nodules (Table I), which is unusual for plant tissues. Two genes encoding Cys synthases were found to be expressed at higher levels in nodules than in roots of Lotus (Colebatch et al., 2004), which could contribute to elevated Cys levels. It is also noteworthy that several genes for sulfate transporters, which presumably deliver substrate for sulfur metabolism, are highly induced during Lotus nodule development (Colebatch et al., 2002, 2004).

While it is not possible to gauge from our GC-MS data the separate contribution that bacteroids make to most metabolite pools, some of the unusual and unidentified compounds that accumulate in nodules, e.g. [802; Methylcitric acid (4TMS)], may be exclusively bacterial products (Table I). Elucidation of the structures of these compounds and their biosynthetic origin will undoubtedly lead to a better understanding of nodule metabolism and the metabolic interactions between legumes and rhizobia. Another important area for future work is metabolic flux determination in nodules. The data presented here give a static picture of metabolite levels averaged over whole organs and provide little insight into metabolic compartmentation or flux though specific pathways. Nonetheless, the resources developed during this project, e.g. MST libraries, will provide a firm basis upon which to build such studies in the future.

MATERIAL AND METHODS

Biological Material, Plant Growth, and Harvesting

Lotus japonicus cv. GIFU seeds were scarified 3 × 10 s in liquid nitrogen, sterilized 10 min in 2% bleach solution, rinsed 5 times with sterile distillated water, and moved to a petri dish with filter paper soaked in B&D medium (Broughton and Dilworth, 1971). After germination in a phytotron set to 25°C and a 16-/8-h day/night cycle, 3-d-old seedlings were transferred to pots, 5 plants each pot, containing coarse quartz sand. Inoculation was performed with *Mesorhizobium loti* strain R7A. Inoculated plants were grown in a greenhouse with a 16-/8-h day/night cycle, 60% relative humidity, a 21°C/17°C day/night temperature regime, and 1 watering/d with B&D medium.

Two sets of experiments were performed. The first set comprised plant material harvested 12 weeks after germination in the course of 1 diurnal cycle, at 2, 8, and 14 h within the light cycle and at 2, 4, and 6 h during the dark period, respectively. While the diurnal changes were not a topic of this investigation, an equal representation of all diurnal stages was generated for a nonbiased organ-to-organ comparison. A second set of experiments was performed in the course of 6 months, early summer to winter. Samples were taken randomly in the middle of the light cycle at different developmental stages, 7 to 17 weeks after germination. Plants were cultivated either in an open pot or a closed glass jar. This experimental set was expected to be highly variable but allowed to verify persistent metabolic features of *L. japonicus* organs. At each harvest, plants were carefully pulled from the quartz sand and a complete set of six organ samples prepared by immediate shock freezing in liquid nitrogen, i.e. nodules, lateral and primary root, mature and developing leaves, and flowers. Leaves were separated according to morphological criteria into a group of young expanding leaves from the apex of the plant and a group of mature fully expanded leaves

from the middle of the plant shoot. Senescent leaves were discarded. Whole flowers were prepared including all floral organs, petals, sepals, carpels, stamen, and pollen. Lateral roots without visible nodule primordia were collected, followed by pink nodules sampled in a representative range of various sizes. The harvest was completed by preparing the primary root, i.e. 2 cm of the main root directly below the hypocotyl. Only samples without nodules and lateral roots were collected. Primary root material had to be sliced before shock freezing to improve subsequent grinding under liquid nitrogen. Samples were stored for a maximum of 4 weeks at −80°C until GC-MS analysis.

GC-MS Metabolite Profiling of Polar Metabolites

Frozen samples of 25 to 50 mg fresh weight were ground for 2 min in 2-mL micro vials with a clean stainless steel metal ball (5-mm diameter) using a ball mill grinder (MM200, Retsch, Haan, Germany) set to 30 cycles/s. All material was thoroughly precooled in liquid nitrogen. Frozen powder was extracted with hot MeOH/$CHCl_3$ and the fraction of polar metabolites prepared by liquid partitioning into water as described earlier (Wagner et al., 2003; Colebatch et al., 2004). Samples were analyzed by GC-MS using a quadrupole type GC-MS system (MD800, ThermoQuest, Manchester, UK). Ribitol, isoascorbic acid, and deuterated Ala were added for internal standardization. Metabolite samples were derivatized by methoxyamination using a 20-mg/mL solution of methoxyamine hydrochloride in pyridine and subsequent trimethylsilylation with N-methyl-N-(trimethylsilyl)-trifluoroacetamide (Fiehn et al., 2000; Roessner et al., 2000). A C_{12}, C_{15}, C_{19}, C_{22}, C_{28}, C_{32}, C_{36} n-alkane mixture was used for the determination of RIs. Details of GC-MS analysis were published previously (Colebatch et al., 2004).

MST Definition and Concept

MSTs are defined as full mass spectra obtained from GC-MS chromatograms. MSTs are described by chromatographic retention, for example RI, and mass spectrum, i.e. a set of fragments that are characterized by *m/z* and relative fragment intensity and normalized to the most abundant fragment of the MS. MSTs represent chemical derivatives of metabolites or metabolites that are not derivatized. MSTs of unidentified compounds can be identified in later experiments by exploiting the above characteristics in standard addition experiments of pure reference substances to the complex biological matrix of interest.

Each mass fragment that belongs to one MST can be used for quantification, and we name these fragments through combination of a single *m/z* and RI from the MST: for example, fragment 279_2758 below. The best fragment for quantification is generally the most abundant one. However, since metabolite profiles comprise hundreds of MSTs of identified and unidentified compounds, mass fragments need to be highly selective. Because metabolite profiles may contain unexpected, novel MSTs, we analyze multiple fragments per MSTs. If all fragments of one MST exhibit the same relative change, we automatically select the most abundant fragment for quantification. If fragment ratios exhibit discrepancies, we manually overrule the automatic choice and select the next best specific fragment for quantification.

MST Identifications and Test for Artifacts

Reference substances for standard addition experiments were from Sigma-Aldrich (Schnelldorf, Germany) except for DL-2-methylcitric acid, which was obtained from C/D/N Isotopes (Pointe-Claire, Quebec, Canada). Lactic acid and benzoic acid were laboratory contaminations as was oligomethylcyclosiloxane, monitored by mass fragment 279_2758. This compound was a chemical artifact caused by the N-methyl-N-(trimethylsilyl)-trifluoroacetamide reagent. Mass spectra were analyzed by AMDIS software (http://chemdata.nist.gov/mass-spc/amdis/; National Institute of Standards and Technology) and compared with commercial and user libraries in NIST02 format (http://chemdata.nist.gov/mass-spc/Srch_v1.7/index.html; National Institute of Standards and Technology). *L. japonicus* MSTs are made available via the Internet at the CSB.DB resource (http://csbdb.mpimp-golm.mpg.de/gmd.html).

Generation of a Metabolite Response Matrix

We manually selected one or more specific mass fragments and corresponding retention time windows for identified and still unidentified MSTs from *L. japonicus*. The find algorithm of the MassLab 1.4v software (ThermoQuest, Manchester, UK) was used to automatically retrieve peak areas and chromatographic retention from GC-MS metabolite profiles. Peak identification and area integration was manually supervised as described above. Peak areas with low intensity were rejected.

In accordance with Colebatch et al. (2004) peak areas, X, were defined to represent the fragment responses (X_i of fragment i). Fragment responses were normalized by fresh weight of the sample and response of the internal standard, ribitol, ($N_i = X_i \times X_{ribitol}^{-1} \times \text{fresh weight}^{-1}$). This procedure corrects pipette errors and slight differences in sample amount. The relative response of a fragment is defined relative to the average normalized response of all tissue samples ($R_i = N_i \times \text{avgN}^{-1}$), namely the average response of flower, nodule, developing leaf, mature leaf, primary root, and lateral root samples.

Statistical Analysis

PCA was performed after $\log_{10}$ transformation of the relative responses, $\log_{10}(R_i)$. Missing values were either manually replaced, in the case of identified MSTs, or defined as average of the respective sample group after $\log_{10}$ transformation. If no response was retrievable for any of the samples of a specific organ, $\log_{10}(R_i) = 0$ was substituted for PCA analysis. HCA was applied to classify MSTs, which represented identified metabolites and a selection of unidentified metabolites, according to their relative abundance in different organs. For this purpose, average normalized responses (avgN_i) were calculated of each MST and organ. Missing data were substituted by the normalized response at detection limit. HCA was performed after range normalization using Euclidian distance and average linkage. All procedures including ANOVA and visualization were performed with EXCEL software and the S-Plus 2000 software package standard edition release 3 (Insightful, Berlin Germany), and multivariate and cluster analysis was essentially as reported earlier (Colebatch et al., 2004).

ACKNOWLEDGMENTS

The authors thank Nicole Gatzke, Cornelia Wagner, and Katrin Bieberich for their patient assistance and technical expertise. The authors greatly appreciate Dr. Andreas Richter (Institute of Ecology and Conservation Biology, Vienna, Austria) for making pinitol and ononitol available for GC-MS standard addition experiments.

Received October 20, 2004; returned for revision December 8, 2004; accepted December 12, 2004.

LITERATURE CITED

Appleby CA (1984) Leghemoglobin and rhizobium respiration. Annu Rev Plant Physiol Plant Mol Biol **35:** 443–478

Ausloos P, Clifton CL, Lias SG, Mikaya AI, Stein SE, Tchekhovskoi DV, Sparkman OD, Zaikin V, Zhu D (1999) The critical evaluation of a comprehensive mass spectral library. J Am Soc Mass Spectrom **10:** 287–299

Batut J, Boistard P (1994) Oxygen control in rhizobium. Antonie Leeuwenhoek J Microbiol Serol **66:** 129–150

Brewin NJ (1991) Development of the legume root nodule. Annu Rev Cell Biol **7:** 191–226

Broughton WJ, Dilworth M (1971) Control of leghemoglobin synthesis in snakes beans. Biochem J **125:** 1075–1080

Chen F, Duran AL, Blount JW, Sumner LW, Dixon RA (2003) Profiling phenolic metabolites in transgenic alfalfa modified in lignin biosynthesis. Phytochemistry **64:** 1013–1021

Colebatch G, Desbrosses G, Ott T, Krusell T, Montanari O, Kloska S, Kopka J, Udvardi MK (2004) Global changes in transcription orchestrate metabolic differentiation during symbiotic nitrogen fixation in Lotus japonicus. Plant J **39:** 487–512

Colebatch G, Kloska S, Trevaskis B, Freund S, Altmann T, Udvardi MK (2002) Novel aspects of symbiotic nitrogen fixation uncovered by transcript profiling with cDNA arrays. Mol Plant Microbe Interact **15:** 411–420

Copeland L, Vella J, Hong ZQ (1989) Enzymes of carbohydrate-metabolism in soybean nodules. Phytochemistry **28:** 57–61
Day DA, Copeland L (1991) Carbon metabolism and compartmentation in nitrogen-fixing legume nodules. Plant Physiol Biochem **29:** 185–201
Doyle JJ, Luckow MA (2003) The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. Plant Physiol **131:** 900–910
Duran AL, Yang J, Wang LJ, Sumner LW (2003) Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). Bioinformatics **19:** 2283–2293
Fedorova M, van de Mortel J, Matsumoto PA, Cho J, Town CD, VandenBosch KA, Gantt JS, Vance CP (2002) Genome-wide identification of nodule-specific transcripts in the model legume Medicago truncatula. Plant Physiol **130:** 519–537
Fernie AR, Tiessen A, Stitt M, Willmitzer L, Geigenberger P (2002) Altered metabolic fluxes result from shifts in metabolite levels in sucrose phosphorylase-expressing potato tubers. Plant Cell Environ **25:** 1219–1232
Fiehn O (2002) Metabolomics: the link between genotypes and phenotypes. Plant Mol Biol **48:** 155–171
Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L (2000) Metabolite profiling for plant functional genomics. Nat Biotechnol **18:** 1157–1161
Fischer HM (1996) Environmental regulation of rhizobial symbiotic nitrogen fixation genes. Trends Microbiol **4:** 317–320
Flemetakis E, Efrose RC, Desbrosses G, Dimou M, Delis C, Aivalakis G, Udvardi MK, Katinakis P (2004) Induction and spatial organization of polyamine biosynthesis during nodule development in Lotus japonicus. Mol Plant Microbe Interact **17:** 1283–1293
Gardiol AE, Truchet GL, Dazzo FB (1987) Requirement of succinate dehydrogenase activity for symbiotic bacteroid differentiation of Rhizobium meliloti in alfalfa nodules. Appl Environ Microbiol **53:** 1947–1950
Gordon AJ, Minchin FR, James CL, Komina O (1999) Sucrose synthase in legume nodules is essential for nitrogen fixation. Plant Physiol **120:** 867–878
Huhman DV, Sumner LW (2002) Metabolic profiling of saponins in Medicago sativa and Medicago truncatula using HPLC coupled to an electrospray ion-trap mass spectrometer. Phytochemistry **59:** 347–360
Joliffe IT (1986) Principal Component Analysis. Springer-Verlag, New York
Karoutis AI, Tyler RT, Slater GP (1992) Analysis of legume oligosaccharides by high-resolution gas-chromatography. J Chromatogr **623:** 186–190
Khalil AH, Eladawy TA (1994) Isolation, identification and toxicity of saponin from different legumes. Food Chem **50:** 197–201
Kopka J, Fernie A, Weckwerth W, Gibon Y, Stitt M (2004) Metabolite profiling in plant biology: platforms and destinations. Genome Biology **5:** 109–117
Matamoros MA, Moran JF, Iturbe-Ormaetxe I, Rubio MC, Becana M (1999) Glutathione and homoglutathione synthesis in legume root nodules. Plant Physiol **121:** 879–888
Miller SS, Driscoll BT, Gregerson RG, Gantt JS, Vance CP (1998) Alfalfa malate dehydrogenase (MDH): molecular cloning and characterization of five different forms reveals a unique nodule-enhanced MDH. Plant J **15:** 173–184
Noiraud N, Maurousset L, Lemoine R (2001) Transport of polyols in higher plants. Plant Physiol Biochem **39:** 717–728
Nuccio ML, Rhodes D, McNeil SD, Hanson AD (1999) Metabolic engineering of plants for osmotic stress resistance. Curr Opin Plant Biol **2:** 128–134
Pathirana SM, Vance CP, Miller SS, Gantt JS (1992) Alfalfa root nodule phosphoenolpyruvate carboxylase: characterization of the cDNA and expression in effective and plant-controlled ineffective nodules. Plant Mol Biol **20:** 437–450
Polhill RM, Raven PH, Stirton CH (1981) Evolution and systematics of the Leguminosae. *In* RM Polhill, PH Raven, eds, Advances in Legume Systematics Part 1. Royal Botanic Gardens, Kew, UK pp 1–26
Robson RL, Postgate JR (1980) Oxygen and hydrogen in biological nitrogen-fixation. Annu Rev Microbiol **34:** 183–207
Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L, Fernie AR (2001a) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. Plant Cell **13:** 11–29
Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L (2000) Technical advance: simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. Plant J **23:** 131–142
Roessner U, Willmitzer L, Fernie AR (2001b) High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. Plant Physiol **127:** 749–764
Roessner-Tunali U, Hegemann B, Lytovchenko A, Carrari F, Bruedigam C, Granot D, Fernie AR (2003) Metabolic profiling of transgenic tomato plants overexpressing hexokinase reveals that the influence of hexose phosphorylation diminishes during fruit development. Plant Physiol **133:** 84–99
Ronson CW, Lyttleton P, Robertson JG (1981) C4-dicarboxylate transport mutants of *rhizobium trifolii* form ineffective nodules on *trifolium repens*. Proc Natl Acad Sci USA **78:** 4284–4288
Roth E, Jeon K, Stacey G (1988) Homology in endosymbiotic systems: the term 'symbiosome'. *In* R Palacios, DPS Verma, eds, Molecular Genetics of Plant-Microbe Interactions. American Phytopathological Society Press, St. Paul, pp 220–225
Sciotti MA, Chanfon A, Hennecke H, Fischer HM (2003) Disparate oxygen responsiveness of two regulatory cascades that control expression of symbiotic genes in Bradyrhizobium japonicum. J Bacteriol **185:** 5639–5642
Steele HL, Werner D, Cooper JE (1999) Flavonoids in seed and root exudates of Lotus pedunculatus and their biotransformation by Mesorhizobium loti. Physiol Plant **107:** 251–258
Stein SE (1999) An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. J Am Soc Mass Spectrom **10:** 770–781
Streeter JG (1980) Carbohydrates in soybean nodules. II. Distribution of compounds in seedlings during the onset of nitrogen fixation. Plant Physiol **66:** 471–476
Streeter JG (1987) Carbohydrate, organic acid and amino acid composition of bacteroids and cytosol from soybean nodules. Plant Physiol **85:** 768–773
Streeter JG, Bosler ME (1976) Carbohydrates in soybean nodules: identification of compounds and possible relationships to nitrogen fixation. Plant Sci Lett **7:** 321–329
Sumner LW, Mendes P, Dixon RA (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era. Phytochemistry **62:** 817–836
Tolstikov VV, Fiehn O (2002) Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. Anal Biochem **301:** 298–307
Tolstikov VV, Lommen A, Nakanishi K, Tanaka N, Fiehn O (2003) Monolithic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics. Anal Chem **75:** 6737–6740
Udvardi MK, Day DA (1997) Metabolite transport across symbiotic membrane of legume nodules. Annu Rev Plant Physiol Plant Mol Biol **48:** 493–523
Udvardi MK, Price GD, Gresshoff PM, Day DA (1988) A dicarboxylate transporter on the peribacteroid membrane of soybean nodules. FEBS Lett **231:** 36–40
Vance CP, Gregerson RG, Robinson DL, Miller SS, Gantt JS (1994) Primary assimilation of nitrogen in alfalfa nodules: molecular-features of the enzymes involved. Plant Sci **101:** 51–64
Vance CP, Reibach PH, Pankhurst CE (1987) Symbiotic properties of Lotus-pedunculatus root-nodules induced by Rhizobium-loti and Bradyrhizobium sp (Lotus). Physiol Plant **69:** 435–442
Wagner C, Sefkow M, Kopka J (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. Phytochemistry **62:** 887–900
Wienkoop S, Saalbach G (2003) Proteome analysis. novel proteins identified at the Peribacteroid membrane from Lotus japonicus root nodules. Plant Physiol **131:** 1080–1090

Physiologia Plantarum 132: 209–219. 2008

REVIEW

Plant metabolomics reveals conserved and divergent metabolic responses to salinity

Diego H. Sanchez[a,†], Mohammad R. Siahpoosh[a,†], Ute Roessner[b], Michael Udvardi[c] and Joachim Kopka[a,*]

[a]Max Planck Institute for Molecular Plant Physiology, Wissenschaftspark Golm, Am Muehlenberg 1, Potsdam-Golm, 14476, Germany
[b]Australian Centre of Plant Functional Genomics, School of Botany, The University of Melbourne, Victoria 3010, Australia
[c]Samuel Roberts Noble Foundation, 2510 Sam Noble Pky., Ardmore, OK 73401, USA

Correspondence
*Corresponding author,
e-mail: kopka@mpimp-golm.mpg.de

Received 28 August 2007; revised 28 September 2007

doi: 10.1111/j.1399-3054.2007.00993.x

New metabolic profiling technologies provide data on a wider range of metabolites than traditional targeted approaches. Metabolomic technologies currently facilitate acquisition of multivariate metabolic data using diverse, mostly hyphenated, chromatographic detection systems, such as GC-MS or liquid chromatography coupled to mass spectrometry, Fourier-transformed infrared spectroscopy or NMR-based methods. Analysis of the resulting data can be performed through a combination of non-supervised and supervised statistical methods, such as independent component analysis and analysis of variance, respectively. These methods reduce the complex data sets to information, which is relevant for the discovery of metabolic markers or for hypothesis-driven, pathway-based analysis. Plant responses to salinity involve changes in the activity of genes and proteins, which invariably lead to changes in plant metabolism. Here, we highlight a selection of recent publications in the salt stress field, and use gas chromatography time-of-flight mass spectrometry profiles of polar fractions from the plant models, *Arabidopsis thaliana*, *Lotus japonicus* and *Oryza sativa* to demonstrate the power of metabolite profiling. We present evidence for conserved and divergent metabolic responses among these three species and conclude that a change in the balance between amino acids and organic acids may be a conserved metabolic response of plants to salt stress.

Introduction

Water limitation is probably the most important environmental constraint affecting global crop productivity. Given the ever-increasing demand for food as the world population rises and the limitation to expand rain-fed crop production, intensive irrigation will be a key strategy in developed and developing countries to meet the global agricultural need. Hence, secondary soil salinization will increasingly affect world agriculture, especially in the expanding arid and semi-arid regions. Unfortunately, we are ill prepared to meet the challenge of soil salinization in agriculture because most traditional crops are salt sensitive. Consequently, intense research effort is now being focused on understanding the physiological basis of salt tolerance in higher plants (Flowers 2004).

New tools for functional genomics have emerged in recent years, including high throughput methods for transcriptomic, proteomic, metabolomic and ionomic

Abbreviations – ANOVA, analysis of variance; FT-IR, Fourier-transformed infrared spectroscopy; ICA, independent component analysis; LC, liquid chromatography; MS, mass spectrometry; PCA, principal component analysis.

†Both authors contributed equally to this work.

analyses, which have significantly enhanced the descriptive power of physiological analysis. These data-rich analytical approaches have sparked development of bioinformatic tools to sift the complex fingerprinting and profiling data sets for relevant descriptive information, which may become important to build predictive models of living organisms or physiological subsystems. More specifically, the metabolite fingerprinting and profiling approaches provide access to the ultimate biological information flow between gene expression and metabolic phenotype.

Because of the highly diverse chemical nature of metabolites, metabolome analyses are subject to technological and analytical constraints, which limit the number of substances that can be accurately identified and quantified from a single sample (e.g. Birkemeyer et al. 2005). Currently, increased comprehensiveness of metabolome coverage can only be achieved by a combination of analytical technologies, which often surpasses the capability of single laboratories. As a result, current publications typically utilize single technological platforms, among which GC-MS, liquid chromatography coupled to mass spectrometry (LC-MS) and nuclear magnetic resonance ^{1}H-NMR are the most common.

The most significant advantage of metabolome analyses is the unchanging chemical identity of metabolite entities. In contrast to genomic, transcriptomic and proteomic analyses where the identity of genes and proteins can be obscured by sequence deviations, metabolomics may prove to be highly applicable to comparative investigations of metabolic phenotypes (Desbrosses et al. 2005), such as the physiological responses caused by environmental perturbations. We highlight a selection of recent publications in the field of metabolomic analyses of the salt-stress response in higher plants and try to provide a brief overview of the current knowledge and applications. Furthermore, we present data from our laboratories providing insights into potentially conserved and divergent metabolic responses to salinity in the plant kingdom and discuss the findings in terms of the current knowledge of salt acclimation physiology.

Metabolite profiling applied to salt-stress physiology

Few studies have utilized metabolic fingerprinting or profiling technologies to discover changes in higher plants upon exposure to salt stress. The metabolic impact of salt stress on crops such as tomato, *Solanum lycopersicon* cv. Edkawy and cv. Simge F1, rice, *Oryza sativa* and grapevine, *Vitis vinifera* cv. Cabernet, and the models *Arabidopsis thaliana* eco. Col-0 as well as *A. thaliana* T87 cell cultures have been investigated (Cramer et al. 2007, Johnson et al. 2003, Kim et al. 2007, Zuther et al. 2007). Comparisons to halophytic species, such as the tree *Populus euphratica* or the shrubs *Thellungiella halophila* and *Limonium latifolium*, have also been performed (Brosche et al. 2005, Gagneul et al. 2007, Gong et al. 2005). Thus, applications range from ecological metabolomic approaches to artificial in vitro systems.

An early study, if not the earliest, utilized Fourier-transformed infrared spectroscopy (FT-IR) for the metabolic fingerprinting of the salt stress response in tomato fruits (Johnson et al. 2003). Several multivariate statistical tools, principal component analysis (PCA), discriminate function analysis and genetic algorithms, provided insight into the salt effect on fruit metabolism by comparison of two tomato varieties with differential impact of salinity with respect to fruit size. Supervised analysis of the FT-IR spectra, which in this case provided information on functional groups or compound classes rather than specific metabolites, revealed that signals from nitrogen containing compounds, particularly nitriles and amino radicals, allowed discrimination of control samples from salt-treated fruits and led to a clear classification of the investigated cultivars, *S. lycopersicon* cv. Edkawy and cv. Simge F1.

An ecophysiological study of the salt-tolerant tree *P. euphratica* combined transcriptomic and GC-MS-based metabolomic analyses (Brosche et al. 2005). This investigation showed that within a natural habitat plants are long-term acclimated to the environment and typically exposed to combinations of environmental stresses. In this study, heat and increased salinity, but not drought stress, were relevant. The authors reported increased amino acid levels, specifically proline, valine and β-alanine, and changes in sugar and polyol metabolism, which appeared to be related to high sodium conditions in the field. The levels of glycerol, glyceric acid and *myo*-inositol increased while those of fructose and mannitol decreased slightly.

The transcriptional and metabolic responses to short-term salt stress were investigated in the glycophyte *A. thaliana* in comparison to the related extremophile shrub *T. halophila*, which combines tolerance to high salinity with drought and freezing resistance (Gong et al. 2005). GC-MS-based metabolite profiling demonstrated interspecies differences, which partially increased in response to salt shock at 150 m*M* NaCl. The most notable result of this study may be the observation that the steady-state pools of many stress-responsive metabolites and transcripts were already changed in *T. halophila* before exposure to salinity, suggesting a constitutive adaptation mechanism in halophytic species. Sugars, e.g. sucrose, fructose and glucose, along with proline and citric, malic and succinic acids were constitutively higher in *T. halophila* than in *A. thaliana*. The raffinose-pathway metabolites,

myo-inositol, galactinol and raffinose accumulated to a greater extent in *T. halophila* than in *A. thaliana* in response to salt stress, while fumaric, malic, phosphoric and aspartic acid levels decreased to a greater extent in the halophyte.

More recently, the in vitro salt-shock response of *A. thaliana* T87 cell cultures was explored by combined GC-MS- and LC-MS-based metabolic profiling (Kim et al. 2007). Early and late intracellular changes of primary metabolism were investigated after exposure to a rapid osmotic perturbation using NaCl. This study revealed short-term and transient induction of the methylation cycle providing methyl groups, the phenylpropanoid pathway for lignin biosynthesis, and of glycine betaine biosynthesis. The long-term response appeared to combine induction of glycolysis and sucrose metabolism.

Cramer et al. (2007) provided a comprehensive comparison between drought and salt stress of pooled shoot tips from grapevine, *V. vinifera* cv. Cabernet. Both transcriptomic and GC-MS-based metabolomic profiling were applied. The latter revealed reduction of sucrose, aspartic, succinic and fumaric acids and the accumulation of proline, asparagine, malic acid and fructose under salt stress. Furthermore, most metabolites exhibited similar trends under water-limited conditions, but glucose, malic acid and proline were more dramatically increased.

The role of the compatible solutes was assessed recently in the halophytic species *L. latifolium*, by means of untargeted and targeted metabolic profiles (Gagneul et al. 2007). Sugars, inositols and proline behaved as osmobalancers, while organic acids decreased upon salt stress. In addition, metabolic responses during acclimation to salinity suggested that many changes in organic solute composition are controlled by constitutive developmental programs, further supporting the hypothesis on a constitutive, adaptive mechanism of halophytes, which may be interpreted as a metabolic anticipation of stress.

Our groups focused on the physiological changes of model glycophytes after long-term salt acclimation to non-lethal salinity levels. We reported the GC-MS analysis of a collection of *O. sativa* cv. grown in hydroponic culture (Zuther et al. 2007). These cultivars were selected to represent a range of differential salt sensitivity and comprised members of both *indica* and *japonica* subspecies. Sensitive *japonica* cv. were clearly distinguished by root and leaf metabolite phenotype from more tolerant cultivars. The sensitive and the most tolerant cultivars were differentiated by metabolic phenotype even under control conditions prior to salt acclimation. Whereas leaf profiles showed only partial distinction of tolerance classes, the root metabolic phenotype provided clearer differentiation and, in addition, separation of the most tolerant cultivars into two classes. A detailed analysis of the root response to salt acclimation showed a strong depletion of tri-carboxylic acids (TCA) cycle intermediates, pyruvic, citric, aconitic, malic and 2-oxoglutaric acids, and of shikimic and quinic acids. This depletion was accompanied by increases in amino acids.

We now extended our investigations to legume salt acclimation physiology, using an integrative ionomics, transcriptomics and metabolomics approach toward a systems analysis of *Lotus japonicus* under different salt doses and experimental acclimation regimes (Sanchez DH, Kopka J, Udvardi M, Model Legumes Congress, 2007, unpublished data) [Correction added after online publication 22 November 2007: This reference has been updated]. Metabolic changes in this model legume were characterized by a general increase in the steady-state levels of many amino acids, sugars and polyols, with a concurrent decrease in concentration of most organic acids. Finally, Roessner et al. (3rd International Conference of the Metabolomics Society, 2007, unpublished data) explored this field by comparing the effects of salinity on the moss, *Physcomitrella patens*, *A. thaliana*, wheat and barley in a number of subspecies/ecotypes that are characterized by differing levels of tolerance. Conserved and species-specific changes of metabolite levels in response to salt stress were observed. In parallel, dramatic differential changes of the metabolic phenotype between tolerant and sensitive cultivars became evident, which appear to be dependent on the species under investigation.

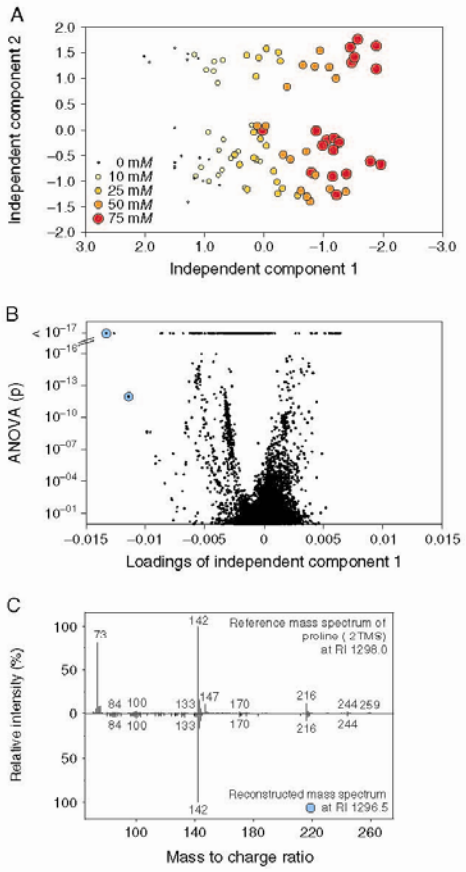
Recent metabolomics-based reports also addressed the changes elicited by desiccation, osmotic or drought stress (e.g. Avelange-Macherel et al. 2006, Cramer et al. 2007, Pinheiro et al. 2004, Rizhsky et al. 2004). Besides investigations of naturally tolerant species and cultivars, the systematic comparison of salt stress and osmotic or drought stress responses may become the upcoming focus of attention. The initial studies appear to be in agreement with the observation that high salinity and desiccation are related, but some distinctions may be found (Cramer et al. 2007).

A general conclusion from metabolomic studies of plant responses to salinity is that changes in metabolism are complex and involve multiple pathways. More specifically, acclimation to salt stress appears to be coupled to changes in organic acid, amino acid and sugar metabolism. While these changes in the pattern of primary metabolism may overlap with responses to other interacting stress factors, a systematic large-scale comparison of different stress factors is still missing.

Experimental design and data mining of GC-MS-based metabolite profiling

Profiling experiments that explore plant stress acclimation typically involve either analysis of time course

dependence or stress dose dependency in order to establish the relevance of physiological observations. Most experiments use randomized arrays of replicate plants and involve relatively few treatments or time points, because of limitations in the numbers of samples that can be processed. Typically 3–6 replications of each treatment are considered a minimal to reasonable requirement for statistical analysis of metabolome experiments. However, the ultimate test for the validity of results is the independent replication of experiments and/or cross-validation procedures. Given the difficulty to reproduce exactly plant experiments in a greenhouse or growth chamber, we started performing at least three repeats of each complete experiment (Fig. 1A).



So far a standardized protocol for the data mining of metabolite profiling data sets is not available. However, widely applied, laboratory-independent methods have emerged, which can be classified into non-supervised and supervised approaches (e.g. Aranibar et al. 2006, Jansen et al. 2004, Johnson et al. 2003, Smilde et al. 2005, Taylor et al. 2002, Trygg et al. 2007). In the following study we applied and combined a non-supervised statistical tool with a supervised method. Non-supervised tools, which find sample classifications within multivariate data sets, are highly suited and perhaps indispensable to estimate the quality of experimentation. PCA and the PCA refinement, independent component analysis (ICA) are two such tools (Daub et al. 2003, Scholz et al. 2004). PCA taps the variance of multivariate data sets and may reveal unexpected sample classifications. These classifications can be described by the so-called principal and independent components (Fig. 1A), which are parameter vectors with defined numerical contributions of each metabolite, the so-called loadings (Fig. 1B). The principal components are sorted by descending variance and thus screening of the first components is often sufficient to identify interesting metabolites. ICA optimizes a selection of the first principal components for best bimodal sample distribution. PCA and ICA can be used to test (1) if the experimental intervention induced a major variance of metabolic phenotype; and (2) visually reveal the relationship of variation within groups of replicate samples compared with variation between sample classes (Fig. 1A). In Fig. 1A, it is important to note that experimental artifact

Fig. 1. Example of non-supervised and supervised analyses of GC-MS metabolite profiles of *A. thaliana* Col-0 shoots acclimated to increasing salt stress doses (0–75 m*M* NaCl). Profiles comprise mass- and chromatography-aligned, $\log_{10}$-transformed response ratios of all mass fragments observed by GC-MS. Response ratios were centered to the control condition (0 m*M* NaCl). (A) Non-supervised analyses, such as ICA, validate expected influences of experimental challenges, e.g. long-term acclimation to salt stress characterized by independent component 1 (IC1), and may reveal non-predictable experimental artifacts, such as differences between independent experiments (IC2). (B) Combinations of non-supervised (ICA) and supervised (ANOVA) analyses allow reduction to the relevant data of profiling experiments. The partial overlap of IC1 loadings analysis and ANOVA alpha-error (p) is demonstrated. Mass fragments exhibiting maximal and minimal loadings values contribute most to the variance induced by the experimental challenge, whereas small p values indicate presence of statistical significance. (C) Mass spectral matching of a reconstructed mass spectrum to reference spectra and retention indices of authenticated metabolite preparations allow unambiguous identification of metabolites, which are relevant for salt acclimation, in this example proline. Mass fragments representing proline are indicated by blue underlay in Fig. 1B. Slight differences between mass fragments representing the same metabolite are caused by varying relative intensities and resulting variations of signal to noise behavior and sensitivity of the mass spectral fragmentation process, which is inherent to GC-MS.

and salt stress are independent factors and that the experimental intervention was a repeatable, common influence in all three experiments. In addition, replicate samples of neighboring salt doses overlap and represent a continuous gradation between control and the highest NaCl dose.

Supervised tools require information on experimental sample classes. The well-known Student's *t*-test checks for differences between two sample classes, whereas analysis of variance (ANOVA) is best applied to compare between multiple sample classes. The so-called post-hoc tests are used to further locate the differences between any of these classes. Both methods are dependent on the number of replications and the distribution properties of measurements. The significance of differences is typically expressed by a probability value or probability threshold (*P*). ANOVA applied to the above data set demonstrates that the influence of variance expressed by ICA loadings and statistical significance may only partially correlate (Fig. 1 B). In other words, the minor variances of a data set may contain valuable discriminatory information. By sheer number or magnitude of multivariate data sets, application of supervised tools bears the risk of generating false-positive results. Therefore, validation by independent biological experiments may be the best practice.

In conclusion, for the task of finding relevant metabolic markers, in this case representing the effects of salt acclimation, a simple workflow for profile analysis may comprise (1) proper experimental design and independent replication; (2) PCA or ICA for non-biased data visualization; (3) ANOVA for reduction of data to those which are relevant for the distinction of sample classes and further secondary PCA or ICA to test for improvement of the expected sample classification; and (4) identification of metabolites from mass fragments with similar statistical properties (Fig. 1C) by mass spectral and retention index matching to reference compounds and spectral libraries (Wagner et al. 2003, Kopka et al. 2005, Schauer et al. 2005, Erban et al. 2007). While more elaborate tools and procedures are available, this work flow was sufficient to select the metabolic responses, which are presented in the following.

Are metabolic responses to salinity conserved or divergent between plant species?

The early reports, but also the new profiling results, summarized above, on the response of higher plants to salt stress appear at first sight to yield variable results. This may be explained in part by inherent differences between plant species, but may also be related to the use of, different organs, time scales of stress exposure, modes of cultivation, e.g. soil and hydroponic growth, and difficult-to-control combinations of stress factors, such as osmotic, drought, salt toxicity, oxidative stress and nutritional

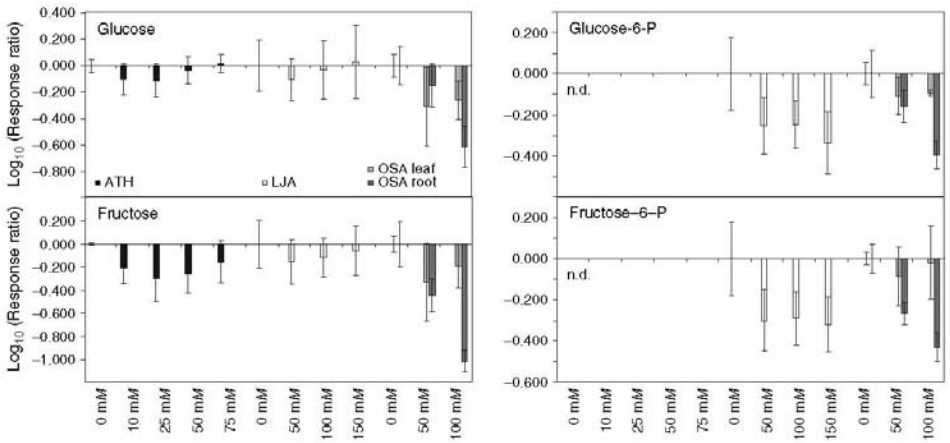


Fig. 2. Dose-dependency pattern of glucose, fructose, glucose-6-P and fructose-6P from salt-acclimated plants. Log_{10}-transformed response ratios representing pool size changes relative to non-treated control plants of each experiment are shown. Error bars represent combined SD of three independent experiments each comprising 4–6 plant replications per condition (e.g. Fig. 1). Shoots of *A. thaliana* Col-0 (ATH), *L. japonicus* var. Gifu (LJA), and both shoots and roots of *O. sativa* ssp. *japonica* (OSA) were submitted to GC-MS metabolite profiling. (n.d. not determined, data below limit of quantification; for method descriptions cf. Appendix S1 in Supplementary Material).

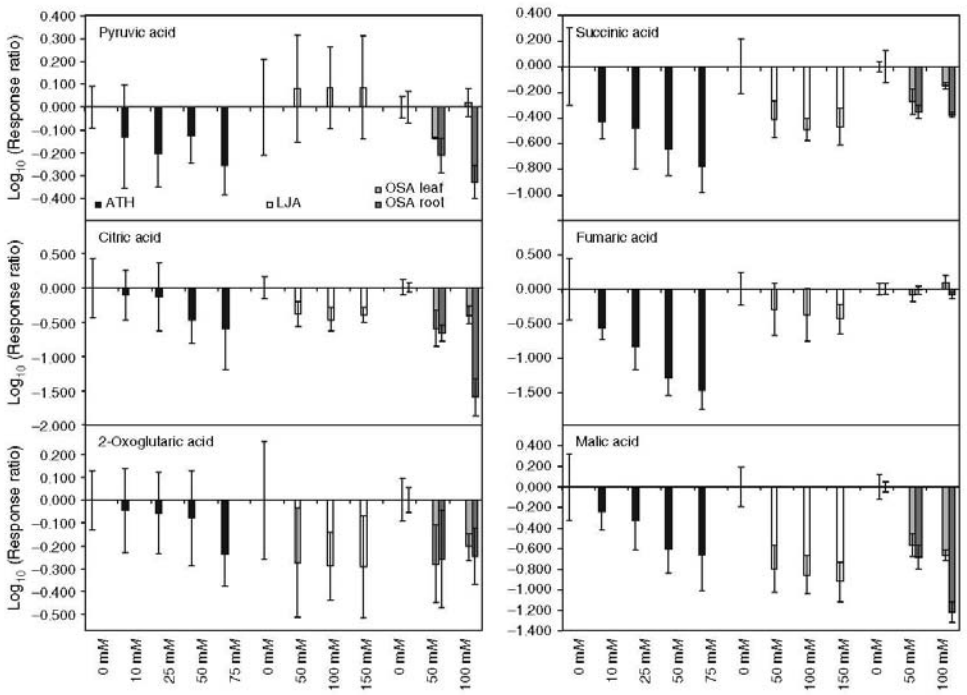


Fig. 3. Dose-dependency pattern of pyruvic, citric, 2-oxoglutaric, succinic, fumaric and malic acids from salt-acclimated plants. (cf. to legend details of Fig. 2).

effects. A thorough systematic reinvestigation appears to be warranted. In the following section, we present comparative gas chromatography time-of-flight mass spectrometry data from *A. thaliana* Col-0, *L. japonicus* var. Gifu, and *O. sativa* ssp. *japonica* (OSA) shoots of plants that were acclimated to non-lethal levels of NaCl (cf. Appendix S1, Methods).

Examples of conserved and divergent metabolic responses to salt acclimation

Conservation of metabolic responses to environmental stress acclimation should be observable within pathways of primary metabolism. In contrast, secondary metabolism, as far as it is not conserved among plants, is more diverse between species and presumably reflects successful adaptation of species through acquisition of novel biosynthetic capacities. GC-MS-based metabolite profiling was initially developed to cover the spectrum of stable polar primary metabolites, such as mono-, di- and tri-saccharides, organic acids, amino acids and amines. Therefore, this technology may be the best method to screen for conservation of metabolic stress responses.

The experiments presented here on salt acclimation of *A. thaliana* Col-0, *L. japonicus* var. Gifu, and OSA were performed under species-specific growth conditions. *A. thaliana* and *L. japonicus* were grown on soil while *O. sativa* was cultivated in hydroponic medium to allow access to root samples. As a result of the use of soil substrate the nitrogen supply was not fully equalized, and, accordingly, the behavior of beta-alanine, glutamine, asparagine and glutamic, aspartic and γ-aminobutyric acids was highly diverse (Fig. S1). Even though the impact of salinity on plant physiology is expected to be strongly influenced not only by nutrient supply, but also by growth condition and physicochemical characteristics of the cultivation substrate (Grattan and Grieve 1999), we found some similarities in the stress-induced changes of steady-state metabolite pools.

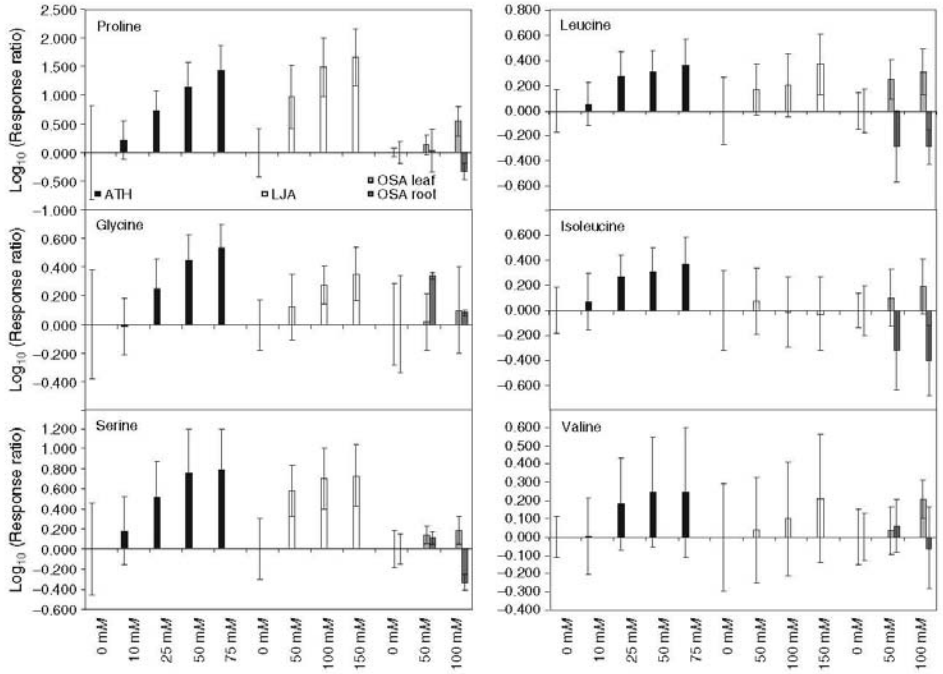


Fig. 4. Dose-dependency pattern of proline, glycine, serine, leucine, isoleucine and valine from salt-acclimated plants. (cf. to legend details of Fig. 2).

Sugar metabolism exhibited conserved, salt dose-dependent accumulation of sucrose. Glucose, fructose and even more so glucose-6-phosphate and fructose-6-phosphate were depleted. Note that both sugar phosphates were below the limit of quantification in the *A. thaliana* experiments. In all three species, glucose and fructose appeared to be decreased at low NaCl concentrations and may revert or even increase at high salt doses (Fig. 2). This conserved reduction of monosaccharide pools appeared to be propagated to acid metabolism. Glyceric, threonic and pyruvic acids were reduced in *A. thaliana* and *O. sativa*. Citric acid, 2-oxoglutaric, succinic and malic acids were strongly reduced in all three species (Fig. 3). In contrast, fumaric acid was reduced in *A. thaliana* and *L. japonicus* but stayed constant in *O. sativa*. Other organic acids such as oxalic and maleic acids, which are not part of the TCA cycle, also exhibited conserved reduction of pool sizes in response to salt-treatment. However, conserved increases were observed among amino acids. Besides the best-known marker of salt acclimation, proline, other amino acids, such as glycine (not increased in *O. sativa* shoot), serine, threonine (data not shown), leucine, isoleucine (not increased in *L. japonicus*) and valine, were increased in a dose-dependent manner (Fig. 4). The changes of these amino acids were generally smaller in *O. sativa* than in the dicotyledonous species.

Metabolic acclimation to stress might be expected to involve species-specific secondary metabolism. Three examples illustrate this point. Galactinol, the biosynthetic precursor of the raffinose-sugar family utilizing *myo*-inositol for transfer reactions of galactose moieties, appears to be a general stress-induced metabolite of *A. thaliana* (Taji et al. 2002). Accordingly, raffinose and galactinol were increased in this species (Fig. 5). However, *O. sativa* exhibited only minor increases of both metabolites, while *L. japonicus* appears to redirect inositol metabolism toward biosynthesis of methyl-inositols, such as ononitol and pinitol. These compounds strongly accumulate in *L. japonicus* under salt stress. Pinitol and ononitol have been suggested to act as compatible solutes in some halophytes (Nelson et al. 1998).

The use of sugar-derived C_6-acids was also highly diverse among species. All three species exhibited galactonic, galactaric and saccharic acid pools above

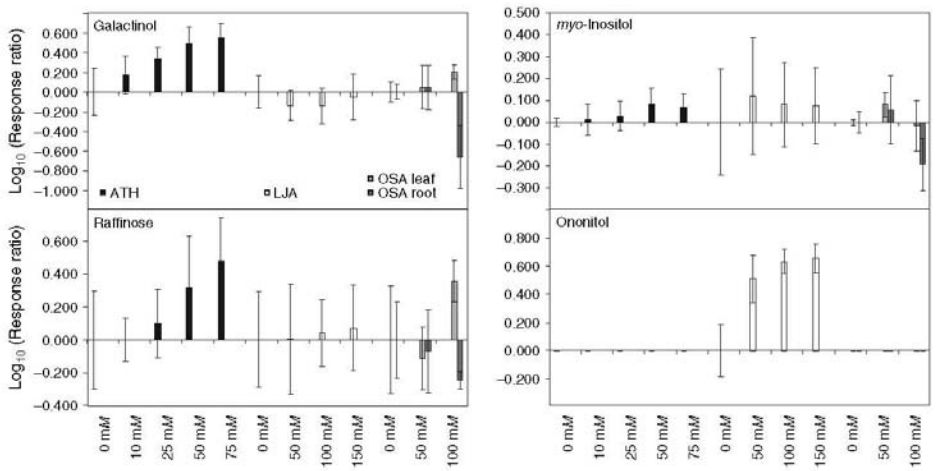


Fig. 5. Dose-dependency pattern of galactinol, raffinose, *myo*-inositol and ononitol from salt-acclimated plants. (cf. to legend details of Fig. 2).

limits of quantification. However, only *O. sativa* had increased levels of these acids in response to salt. In contrast, *A. thaliana* specifically accumulated gluconic acid and showed a slight decrease of galactonic and galactaric acids. Moreover, *L. japonicus* exhibited increased gulonic and glucuronic acid levels (Fig. S2).

The final example is taken from the amine and amino acid metabolism of *O. sativa*. In response to salt stress, the leaves of this species accumulate tryptophan, 5-hydroxy-tryptamine and to a minor degree the intermediate tryptamine. This response was not observed in the dicotyledonous plants (Fig. 6). This example also shows that acclimation of root and shoot metabolisms may differ substantially. Tryptophan, tryptamine and 5-hydroxy-tryptamine show reduction in roots under salt acclimation, similar to leucine, isoleucine, galactinol and raffinose. However, some responses such as the changes of organic acids and glucose or fructose appear to occur in both shoot and root organs.

The major metabolic patterns of salt acclimation

The metabolic window, which can be accessed by the current profiling technologies, is biased by method of extraction and quantification. The coverage of metabolism through GC-MS profiling is still narrow and includes primary metabolism and other metabolites that are available as pure authenticated reference compounds. However, the widened scope of metabolic coverage compared with traditional targeted analyses may lead to hypotheses on general patterns of the metabolic phenotypes under salt acclimation. The observed metabolic changes may represent general or salt-stress-specific responses. Elucidation of pattern specificity by systematic comparison of environmental stress factors and extension of experiments from dose dependency to time course behavior will be task of the future.

Salinity imposes both osmotic and ionic stress components on plants. This impact of salinity on plant physiology is described in the so-called biphasic growth model (Munns 2002, 2005). According to this model, a sudden increase in soil salinity will limit plant growth in two phases. The first phase is caused by decrease of water potential, which is dependent on salt concentration and therefore should be similar to osmotic or drought stress. The second phase is triggered by accumulation of ions within plant tissues, which affects metabolism and nutrient availability. Different pathways might be elicited in sequence by reduced water availability and/or ionic imbalance. Therefore, at least two phases of metabolic changes upon salt stress may be predicted, with the first phase showing similarities to osmotic or drought stress.

Published results from salt-shock or short-term exposure to high salt concentrations suggest a substantial similarity between salt and drought responses. Increases in some primary metabolites were reported, which are typically linked to amino acid and nitrogen metabolism,

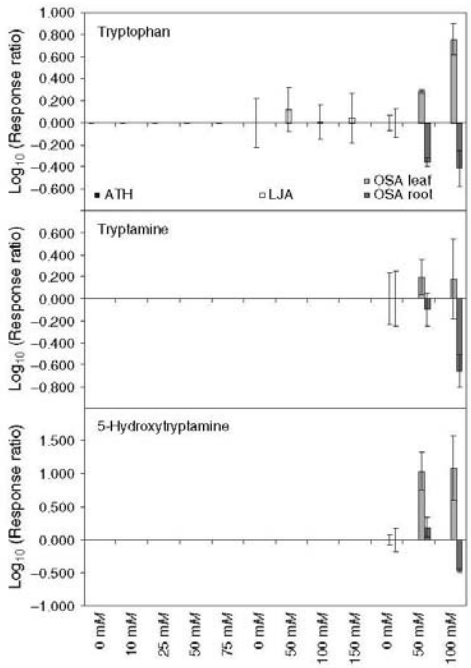


Fig. 6. Dose-dependency pattern of tryptophan, tryptamine and 5-hydroxytryptamine from salt-acclimated plants. (cf. to legend details of Fig. 2).

or carbohydrate and polyol metabolism (e.g. Cramer et al. 2007, Kim et al. 2007, Rizhsky et al. 2004). Such metabolites are often considered to be compatible solutes, which serve osmotic adjustment, protect membranes and proteins, scavenge free oxygen radicals and act as repository of carbon and nitrogen (Bohnert et al. 1995). A strong argument has been made for the role of nitrogen containing compounds, acting as metabolic scavengers of excess ammonium accumulated under stress (Rabie, 1999). The content of nitrogen intermediates, however, is highly influenced by nitrogen nutrition.

In agreement with their suspected role as compatible solutes and osmoprotectants, the steady-state pools of nitrogen-containing compounds and several carbohydrates and polyols are not only modified by fast osmotic changes but are also highly elicited under extended salt treatments (e.g. Cramer et al. 2007, Zuther et al. 2007). Salt-acclimated plants appear to exhibit a second remarkable metabolic feature, namely most organic acid pools are depleted (e.g. Gong et al. 2005, Zuther et al. 2007). Depletion of organic acid intermediates may be interpreted as a general consequence of decreased de novo assimilation of carbon dioxide under stress because of stomatal limitation. However, osmotic stress, drought, cold and heat appear to be accompanied by an accumulation of organic acid pools (e.g. Kaplan et al. 2004, Timpa et al. 1986). Moreover, a similar feature was described in halophytic species, suggesting that this is a metabolically conserved response not related to salt sensitivity (Gagneul et al. 2007, Gong et al. 2005). A function of reduced acid levels under salt stress may be the involvement in the compensation of ionic imbalance, because at physiological pH organic acids exist as carboxylic anions and can therefore counterbalance inorganic anions. A depletion of organic acid levels may actually reflect preferential uptake of anions compared with cations (Hinsinger et al. 2003, Marschner 1995), which may be concomitant with the decrease of the theoretically balancing anions that occurs under long-term salt-stressed plants (Munns et al. 2006). As a consequence, the reduction of carboxylic acids levels may be prerequisite for maintaining vitality. Taken together, the depletion of organic acids may serve several functions. Surplus organic acids may indeed be recruited from the TCA cycle and sequestered into biosynthesis pathways of amino acids and amines. Thus all demands, (1) maintenance of charge balance; (2) ammonium detoxification; and (3) compatible solute accumulation, may be met.

Concluding remarks

We argue that combinatorial interactions between the ion balance, mineral nutrition and growth responses under salt stress are reflected in controlled conserved and divergent changes of primary and secondary metabolisms during salt stress in plants. These changes induce a massive change of the metabolic phenotype in response to salt acclimation. In agreement with the biphasic model of salt response, current metabolic data indicate similarity of early changes with osmotic or drought responses while long-term salt acclimation may reflect the attempt to cope with increasing ion toxicity finally reaching a critical state dependent on dose and time of salt exposure. Metabolomic profiling provides insight into features of stress-induced changes in higher plants, revealing not only conserved but also species-specific responses during salt acclimation. To substantiate hypotheses on beneficial metabolic states or biosynthesis capabilities, systematic investigation of model plants and crop species is clearly required. In the future, detailed analyses of the effects of salt acclimation and related stress factors in time and dose may help to direct crop breeding programs toward increased salt tolerance.

Acknowledgements – This review was made possible by funding through the European Union LOTASSA project (INCO-CT-2005-517617) and the German Bundesministerium für Bildung und Forschung (BMBF), grant PTJ-BIO/0312854. The data processing and mining was supported by the European META-PHOR project (FOOD-CT-2006-036220). We would like to acknowledge the long-standing and steady support of all directors at the Max Planck Institute for Molecular Plant Physiology (MPIMP), the technical assistance of Ines Fehrle and Alexander Erban by performing GC/EI-TOF-MS profiling analyses and the expert support of the MPIMP greenhouse team, headed by Dr Karin Köhl. U. R. thanks the Australian Centre for Plant Functional Genomics for funding.

References

Aranibar N, Ott KH, Roongta V, Mueller L (2006) Metabolomic analysis using optimized NMR and statistical methods. Anal Biochem 355: 62–70

Avelange-Macherel MH, Ly-Vu B, Delauna Y, Richomme P, Leprince O (2006) NMR metabolite profiling analysis reveals changes in phospholipid metabolism associated with the re-establishment of desiccation tolerance upon osmotic stress in germinated radicles of cucumber. Plant Cell Environ 29: 471–482

Birkemeyer C, Luedemann A, Wagner C, Erban A, Kopka J (2005) Metabolome analysis: the potential of in vivo labeling with stable isotopes for metabolite profiling. Trend Plant Sci 23: 28–33

Bohnert HJ, Nelson DE, Jensen RG (1995) Adaptations to environmental stresses. Plant Cell 7: 1099–1111

Brosche M, Vinocur B, Alatalo ER, Lamminmaki A, Teichmann T, Otow EA, Djilianov D, Afif D, Bogeat-Triboulot M, Altman A, Dreyer E, Rudd S, Paulin L, Auvinen P, Kangasjarvi J (2005) Gene expression and metabolite profiling of *Populus euphratica* growing in the Negev desert. Genome Biol 6: R101

Cramer GR, Ergul A, Grimplet J, Tillett RL, Tattersall EAR, Bohlman MC, Vincent D, Sonderegger J, Evans J, Osborne C, Quilici D, Schlauch KA, Schooley DA, Cushman JC (2007) Water and salinity stress in grapevines: early and late changes in transcript and metabolite profiles. Funct Integr Genomics 7: 111–134

Daub CO, Kloska S, Selbig J (2003) MetaGeneAlyse: analysis of integrated transcriptional and metabolite data. Bioinformatics 19: 2332–2333

Desbrosses GG, Kopka J, Udvardi MK (2005) *Lotus japonicus* metabolic profiling. Development of gas chromatography-mass spectrometry resources for the study of plant-microbe interactions. Plant Physiol 137: 1302–1318

Erban A, Schauer N, Fernie AR, Kopka J (2007) Non-supervised construction and application of mass spectral and retention time index libraries from time-of-flight GC-MS metabolite profiles. In: Weckwerth W (ed) Metabolomics: Methods and Protocols. Humana Press, Totowa, pp 19–38

Flowers TJ (2004) Improving crop salt tolerance. J Exp Bot 55: 307–319

Gagneul D, Ainouche A, Duhaze C, Lugan R, Lahrer FR, Bouchereau A (2007) A reassessment of the function of the so-called compatible solutes in the halophytic Plumbginaceae *Limonium latifolium*. Plant Physiol 144: 1598–1611

Gong Q, Li P, Ma S, Rupassara SI, Bohnert H (2005) Salinity stress adaptation competence in the exptremophile *Thellungiella halophila* in comparison with its relative *Arabidopsis thaliana*. Plant Physiol 44: 826–839

Grattan SR, Grieve CM (1999) Mineral nutrient acquisition and response by plants grown in saline environments. In: Pessarakli M (ed) Handbook of Plant and Crop Stress, 2nd Edn. Marcel Dekker Inc., New York, pp 203–229

Hinsinger P, Plassard C, Tang C, Jaillard B (2003) Origins of root-mediated pH changes in the rhizosphere and their responses to environmental constraints: a review. Plant Soil 248: 43–59

Jansen JJ, Hoefsloot HCJ, Boelens HFM, Van der Greef J, Smilde AK (2004) Analysis of longitudinal metabolomics data. Bioinformatics 20: 2438–2446

Johnson HE, Broadhurst D, Goodacre R, Smith AR (2003) Metabolic fingerprinting of salt stressed tomatoes. Phytochemistry 62: 919–928

Kaplan F, Kopka J, Haskell DW, Zhao W, Schiller KC, Gatzke N, Sung DY, Guy CL (2004) Exploring the temperature-stress metabolome of Arabidopsis. Plant Physiol 136: 4159–4168

Kim JK, Bamba T, Harada K, Fukusaki E, Kobayashi A (2007) Time-course metabolic profiling in *Arabidopsis thaliana* cell cultures after salt stress treatment. J Exp Bot 58: 415–424

Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmuller E, Dormann P, Weckwerth W, Gibon Y, Stitt M, Willmitzer L, Fernie AR, Steinhauser D (2005) GMD@CSB.DB: the Golm Metabolome Database. Bioinformatics 21: 1635–1638

Marschner H (1995) Mineral Nutrition of Higher Plants, 2nd Edn. Academic Press Limited, London

Munns R (2002) Comparative physiology of salt and water stress. Plant Cell Environ 25: 239–250

Munns R (2005) Genes and salt tolerance: bringing them together. New Phytol 167: 645–663

Munns R, James RA, Lauchi A (2006) Approaches to increasing the salt tolerance of wheat and other cereals. J Exp Bot 57: 1025–1043

Nelson DE, Rammesmayer G, Bohnert HJ (1998) Regulation of cell-specific inositol metabolism and transport in plant salinity tolerance. Plant Cell 10: 753–764

Pinheiro C, Passarinho JA, Pinto Ricardo C (2004) Effect of drought and rewatering on the metabolism of

Lupinus albus organs. J Plant Physiol 161: 1203–1210
Rabie E (1999) Altered nitrogen metabolism under environmental stress conditions. In: Pessarakli M (ed) Handbook of Plant and Crop Stress, 2nd Edn. Marcel Dekker Inc, New York, pp 349–363
Rizhsky L, Liang HJ, Shuman J, Shulaev V, Davletova S, Mittler R (2004) When defence pathways collide. The response of *Arabidopsis* to a combination of drought and heat stress. Plant Physiol 134: 1683–1696
Schauer N, Steinhauser D, Strelkov S, Schomburg D, Allison G, Moritz T, Lundgren K, Roessner-Tunali U, Forbes MG, Willmitzer L, Fernie AR, Kopka J (2005) GC-MS libraries for the rapid identification of metabolites in complex biological samples. FEBS Lett 579: 1332–1337
Scholz M, Gatzek S, Sterling A, Fiehn O, Selbig J (2004) Metabolite fingerprinting: detecting biological features by independent component analysis. Bioinformatics 20: 2447–2454
Smilde AK, Jansen JJ, Hoefsloot HCJ, Lamers AN, Ver der Greef J, Timmerman ME (2005) ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. Bioinformatics 21: 3043–3048
Taji T, Ohsumi C, Iuchi S, Seki M, Kasuga M, Kobayashi M, Yamaguchi-Shinozaki K, Shinozaki K (2002) Important roles of drought- and cold- inducible genes for galactinol synthase in stress tolerance in *Arabidopsis thaliana*. Plant J 29: 417–426
Taylor J, King RS, Altmann T, Fiehn O (2002) Application of metabolomics to plant genotype discrimination using statistics and machine learning. Bioinformatics 18: S241–S248
Timpa JD, Burke JJ, Quisenberry JE, Wendt CW (1986) Effects of water-stress on the organic acid and carbohydrate compositions of cotton plants. Plant Physiol 82: 724–728
Trygg J, Holmes E, Lundstedt T (2007) Chemometrics in metabonomics. J Proteome Res 6: 469–479
Wagner C, Sefkow M, Kopka J (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. Phytochemistry 62: 887–900
Zuther E, Koehl K, Kopka J (2007) Comparative metabolome analysis of the salt response in breeding cultivars of rice. In: Jenks MA, Hasegawa PM, Jain SM (eds) Advances in Molecular Breeding Toward Drought and Salt Tolerant Crops. Springer-Verlag, Berlin, Heidelberg, New York, pp 285–315

Supplementary material

The following supplementary material is available for this article:

Appendix S1. Supplemental methods.

Fig. S1. Dose-dependency pattern of β-alanine, glutamine, asparagine and aspartic, glutamic and γ-aminobutyric acids in salt-acclimated plants.

Fig. S2. Dose-dependency pattern of galactonic, gluconic, gulonic, galactaric, glucaric and glucuronic acids from salt-acclimated plants.

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1399-3054.2007.00993.x

(This link will take you to the article abstract).

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Edited by C. Guy

The Plant Journal (2008) **53**, 973–987 doi: 10.1111/j.1365-313X.2007.03381.x

Integrative functional genomics of salt acclimatization in the model legume *Lotus japonicus*

Diego H. Sanchez[1], Felix Lippold[1], Henning Redestig[1], Matthew A. Hannah[1], Alexander Erban[1], Ute Krämer[2], Joachim Kopka[1] and Michael K. Udvardi[3,*]

[1]*Max Planck Institute for Molecular Plant Physiology, Wissenschaftspark Golm, Am Mühlenberg 1, Potsdam-Golm, D-14 476, Germany,*

[2]*Bioquant Center, University of Heidelberg, Im Neuenheimer Feld 267, D-69120 Heidelberg, Germany, and*

[3]*Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73 401, USA*

Received 25 September 2007; accepted 19 November 2007.

*For correspondence (fax +1 580 224 6692; e-mail mudvardi@noble.org).

Summary

The model legume *Lotus japonicus* was subjected to non-lethal long-term salinity and profiled at the ionomic, transcriptomic and metabolomic levels. Two experimental designs with various stress doses were tested: a gradual step acclimatization and an initial acclimatization approach. Ionomic profiling by inductively coupled plasma/atomic emission spectrometry (ICP-AES) revealed salt stress-induced reductions in potassium, phosphorus, sulphur, zinc and molybdenum. Microarray profiling using the Lotus Genechip® allowed the identification of 912 probesets that were differentially expressed under the acclimatization regimes. Gas chromatography/mass spectrometry-based metabolite profiling identified 147 differentially accumulated soluble metabolites, indicating a change in metabolic phenotype upon salt acclimatization. Metabolic changes were characterized by a general increase in the steady-state levels of many amino acids, sugars and polyols, with a concurrent decrease in most organic acids. Transcript and metabolite changes exhibited a stress dose-dependent response within the range of NaCl concentrations used, although threshold and plateau behaviours were also observed. The combined observations suggest a successive and increasingly global requirement for the reprogramming of gene expression and metabolic pathways to maintain ionic and osmotic homeostasis. A simple qualitative model is proposed to explain the systems behaviour of plants during salt acclimatization.

Keywords: acclimatization, ionomic, *Lotus*, metabolomic, salt stress, transcriptomic.

Introduction

Plant salt stress has become a major concern worldwide due to the salinization of agricultural land caused by irrigation, so-called secondary salinization. This is of particular relevance as most crops are salt-sensitive and progressing desertification imposes a need for irrigation. Salinity imposes at least two primary stresses on plants: hyperosmotic stress caused by the reduction of water potential and consequently reduced water availability, and hyperionic stress, related to the toxic effects of the accumulated ions. Consequently, salinized plants are subjected to dehydration, metabolic toxicity, nutrient deficiencies, membrane dysfunction and oxidative stress, which lead to tissue damage and early senescence (Tester and Davenport, 2003). However, plants under salt stress do not show symptoms of cellular damage provided that the stress dose is not prolonged or is below the tolerance threshold. As sessile organisms, plants have evolved a number of strategies to acclimatize to various kinds of deleterious conditions and thus to increase competitiveness in various ecological niches. Unlike adaptation, which is a consequence of evolutionary mechanisms acting at the genetic level in populations over many generations, acclimatization is a proximal phenotypic response to changes in the environment (Orcutt and Nilsen, 2000). Over seconds, minutes, hours or days, plants can reprogram their metabolism, physiology and morphology to attain new physiological states, which ensures fitness and survival under abiotic and biotic constraints (Lichtenthaler, 1996). Among others, salinity acclimatization responses include: (i) the maintenance of ion homeostasis, including ion exclusion, compartmentation,

redistribution, organ-specific allocation and excretion; (ii) osmotic adjustment and compatible solute accumulation; (iii) water balance and control of transpiration; and (iv) structural and anatomical changes, for example the modification of apoplastic barriers (Hose *et al.*, 2001; Tester and Davenport, 2003). Recent evidence also demonstrates that growth and developmental responses are fundamental to plant survival under prolonged salt stress (Achard *et al.*, 2006).

Despite knowledge that plants respond progressively to hyperosmotic and hyperionic stress over time, the mainstream of current molecular and biochemical research is focused on short-term signalling and early responses triggered by salts, and experiments are typically performed using lethal treatments (Munns, 2005). As a result, detailed knowledge of the molecular basis of salt-stress acclimatization is still largely lacking.

Legumes are second in importance to agriculture after grasses, and cover around 12–15% of the world's agricultural land and supply 33% of human dietary nitrogen needs (Graham and Vance, 2003). They also play a critical role in natural and agricultural ecosystems, due to their ability to fix nitrogen. Given their importance, further knowledge of legume stress physiology is required to address present and future threats to food security. Here we present ionomic, transcriptomic and metabolomic analyses of the glycophyte model legume *Lotus japonicus* subjected to long-term regimes of non-lethal levels of salinity. The existence of perennial cultivated *Lotus* species that are particularly salt-tolerant provides a reason for using this model (Teakle *et al.*, 2006). The results presented below reveal new molecular and metabolic components of the salt-stress response in legumes, and provide systems-level insights into the plastic acclimation process. Finally, a simple model is proposed that incorporates these observations to qualitatively explain the dose dependency, plateau and threshold behaviour of the responses.

Results

Experimental set-up and physiological assessment of salt-acclimatized plants

To address the responses of plants during salt acclimatization, two long-term experimental designs were used. The first regime was based on a gradual acclimatization to salts, whereas the second was an initial acclimatization (ia) approach involving seed germination and growth on a range of defined salt concentrations (Figure 1a). Three independent experiments were performed in a greenhouse, each comprising controls and six treatments. These were labelled according to experimental design and final NaCl concentration as 50, 100 and 150 (gradual acclimatization), and ia25, ia50 and ia75 (initial acclimatization).

As expected, the final shoot biomass of salt-acclimatized plants decreased with increasing NaCl concentration (Figure 1b). Shoot Na^+ content correlated linearly with increasing levels of salt added (r^2 = 0.9418 and 0.9471 for the gradual and initial regimes, respectively), but the slope of the regression differed depending on the experimental approach (Figure 1c). In addition, soil conductivity was also linearly correlated (r^2 = 0.9523) with the amount of NaCl added, indicating that salts did not accumulate differentially in the soil due to experimental design or the varying transpiration rates of plants (Figure 1d). As a consequence, the basic difference between the two approaches was that plants that were gradually acclimatized faced a higher level of osmotic stress in the roots for a given internal Na^+ accumulation than those that were initially acclimatized, while the latter were under higher ionic stress than the former at the same final soil salt content.

Nutrient profiling

Inductively coupled plasma/atomic emission spectrometry (ICP-AES) was used to profile changes in shoot micro- and macronutrient contents (Table S1). ANOVA analysis was performed to identify elements with altered levels associated with salt stress, using a false discovery rate lower than 1% (FDR < 0.01, Figure 2). As expected, Na^+ and K^+ levels were negatively correlated under increasing salinity. Calcium and magnesium were slightly increased under all salt treatments and manganese in some, to not more than 150% of controls. The strongest salt-induced decrease was observed for molybdenum, reaching approximately 30% of the control content. Sulphur and phosphorus decreased compared to the control to a minimum of 60% under the ia75 treatment. In addition, zinc decreased only under the most extreme doses of the initial acclimatization design, reaching around 70% of the control content. Iron and boron levels were not significantly altered by any of the salt treatments, and the contents of cadmium, cobalt, chromium, copper, nickel and selenium were below the detection limits.

Gene expression analysis

Transcriptomic analysis using the Affymetrix Lotus Genechip® was performed in three independently replicated experiments to identify genes that were differentially expressed during salt acclimatization. Gene expression was analysed after hierarchical clustering/supporting tree (HCL-ST) analysis applied to the complete set of probesets. Treatments did not cluster either according to the experimental design or increasing soil conductivity. Rather, a trend coincident with shoot Na^+ accumulation was observed (Figure 3a, compare with Figure 1c,d). The data set was further analysed by a significance-based comparison of salt-treated plants and controls, applying an FDR < 0.01 and a

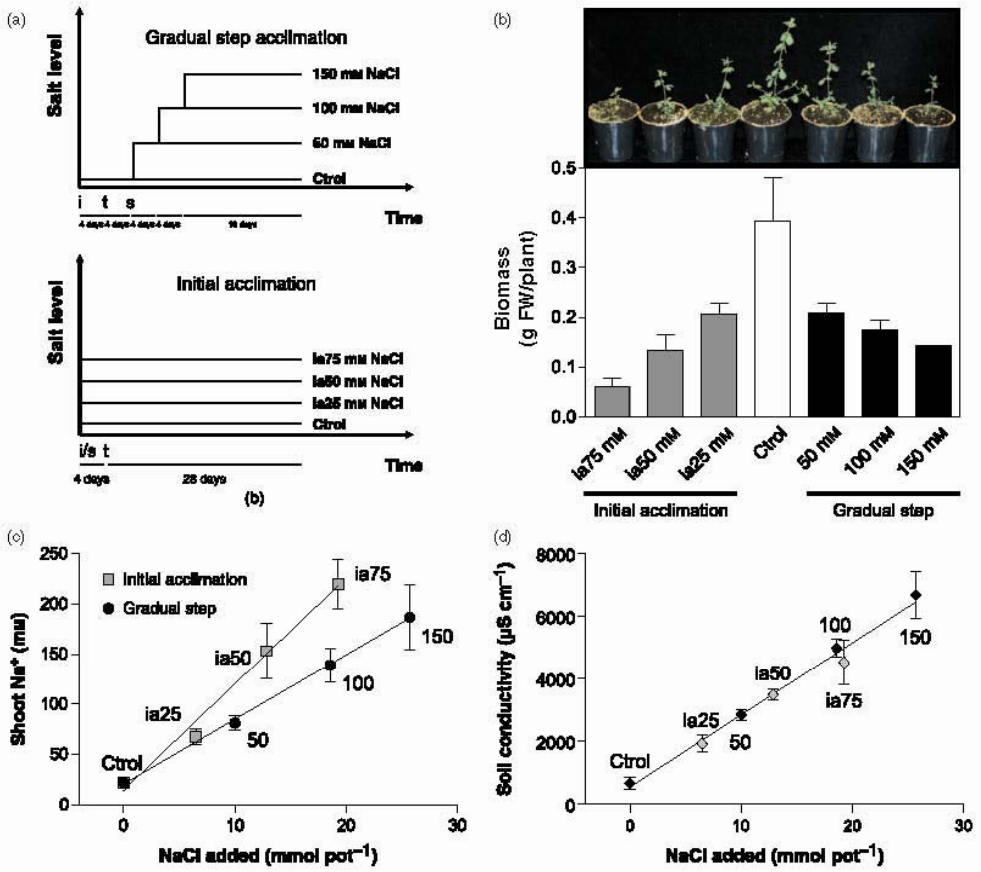


Figure 1. Experimental design and physiological assessment of salt-acclimatized plants.
(a) Gradual step acclimatization and initial acclimatization (ia) experimental designs (see Experimental procedures). i, imbibition; t, transplanting; s, salinization; d, days.
(b) Plant growth and final shoot biomass.
(c) Shoot sodium content.
For (b) and (c), data are the mean ± SD of three independent experiments. Plants photographs were taken at the end of the experiment after 28 days under greenhouse conditions. FW, fresh weight.
(d) Soil conductivity. Data are the mean ± SD of three independent experiments, each containing three independent replicated samples taken from pooled soil from five pots.

twofold change threshold. An increasing number of differentially regulated genes resulted from increasing salinity under both experimental designs (Figure 3b). A total of 912 probesets that matched both statistical and threshold criteria were classified as salt-stress-responsive (522 up-regulated and 390 down-regulated, Table S2). In agreement with the concept of increasing salt toxicity within shoot tissue, treatments with the highest salt accumulation showed an increased number of differentially regulated genes.

As transcription factors are often expressed at low levels and may be difficult to quantify by microarrays (Czechowski *et al.*, 2004), quantitative real-time RT-PCR was used to validate array data for 40 probesets representing putative *L. japonicus* transcription factors. Genes classified or not as salt-regulated were tested, and the mean expression ratio between the 150 mM NaCl treatment and control samples was calculated from both chip and quantitative real-time RT-PCR data for the three independent experiments (Figure 3c

 The Plant Journal, (2008), 53, 973–987

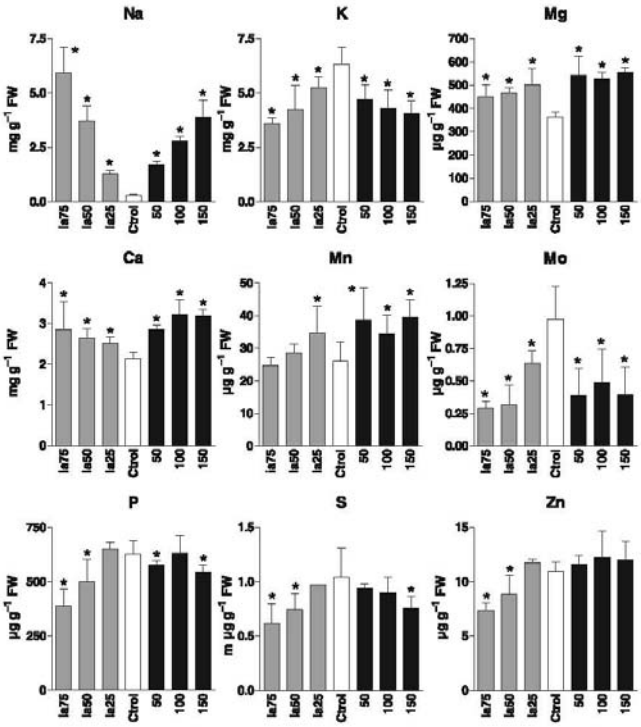
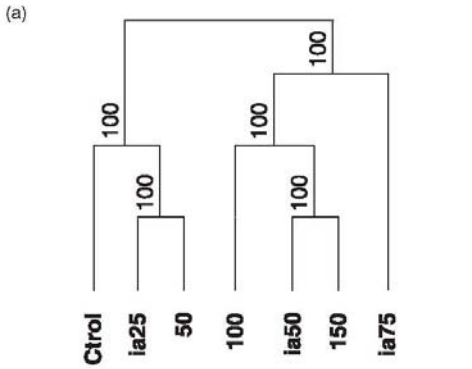


Figure 2. Shoot nutrients exhibiting statistically significant changes compared to the control treatment.
Bars indicate control (white), gradual step acclimatization treatment (black) and initial acclimatization treatment (grey). Data are the mean ± SD of three independent experiments. Asterisks indicate statistically significant changes compared to the control at FDR < 0.01.

and Table S3). Remarkably good agreement was found between the two technologies for 39 probesets, with a linear regression slope of 0.9989 and $r^2 = 0.7976$. A single probeset failed the validation test (chr2.CM0249.113); it was strongly down-regulated according to the chip analysis but not changed according to the quantitative real-time RT-PCR data.

The expression pattern of the 50 most up-regulated and 50 most down-regulated probesets is shown in Figure 4(a). Dose-dependent induction or repression of these genes was evident, irrespective of the experimental approach. Remarkably, a subset of genes deviated from the general trend, responding only at high salt concentrations, while others appeared to reach a plateau of high or low expression. Representative expression patterns are shown in Figure 4(b) for genes exemplifying these behaviours. Functional annotation of salt-regulated probesets was obtained by comparison of all translated sequences to the *Arabidopsis thaliana* genome, and 807 probesets (88%) had a significant hit (E value ≤1e-5). These identifications were used to visualize the functional classification using MapMan software (Usadel *et al.*, 2005). After manual checking, the most common functional groups were: transcription and RNA processing (10%), large enzyme families (miscellaneous, 10%), transport (8%), protein modification and degradation (7%), signalling (7%), stress and defence (6%), cell wall (5%) and hormone metabolism (4%), secondary metabolism (4%) and amino acid metabolism (2%) (Figure 4c). All categories comprised up- and down-regulated probesets, with the exception of the amino acid and cell-wall-related genes. Most of the probesets were up-regulated in the former and down-regulated in the latter, suggesting a potential requirement for salt-stress-specific regulation of these metabolic pathways (Figure S1). Subsequently, we focused on particular functional groups, for which expression of selected target genes was validated with quantitative real-time RT-PCR. Their expression was also tested in whole shoots subjected to salt acclimatization and drought, in seedlings under salt shock, and in various aerial organs under control and salt-acclimatization conditions.

Transcription and RNA processing-related genes. Many transcription factors were found within this group, which presumably coordinate global transcriptional changes during the salt-acclimatization response. All transcription factor hits were verified at http://daft.cbi.pku.edu.cn/



(b)

	ia75	ia50	ia25	Ctrol	50	100	150
FDR < 0.01	1228	813	14	x	49	386	925
> 2 fold	1496	398	16	x	90	432	1007
intersection	767	247	0	x	27	190	506

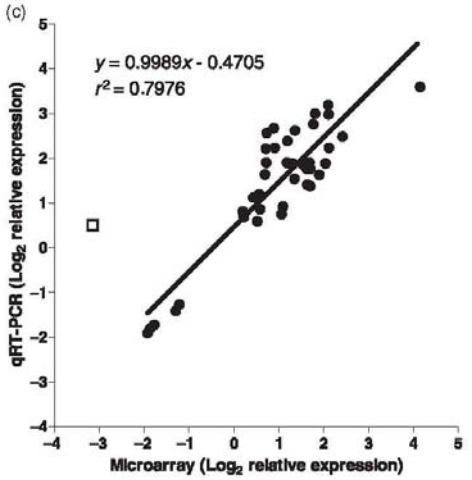


Figure 3. Transcriptomic fingerprinting and microarray results and validation.
(a) Hierarchical clustering/supporting tree analysis (HCL-ST) of the transcriptomic profiles. Bootstrap analysis comprised 10 000 iterations.
(b) Number of probesets from microarray profiling declared salt-responsive in each treatment, based on the false discovery rate (FDR < 0.01), fold change (> twofold) and intersection of the thresholds.
(c) Comparison of microarray and quantitative real-time RT-PCR data from 40 probesets coding for putative *L. japonicus* transcription factors. Each symbol represents the mean expression level (log_2-transformed) of the 150 mM NaCl treatment relative to control treatment from three independent experiments. The open square represents the probeset chr2.CM0249.113 (see Results).

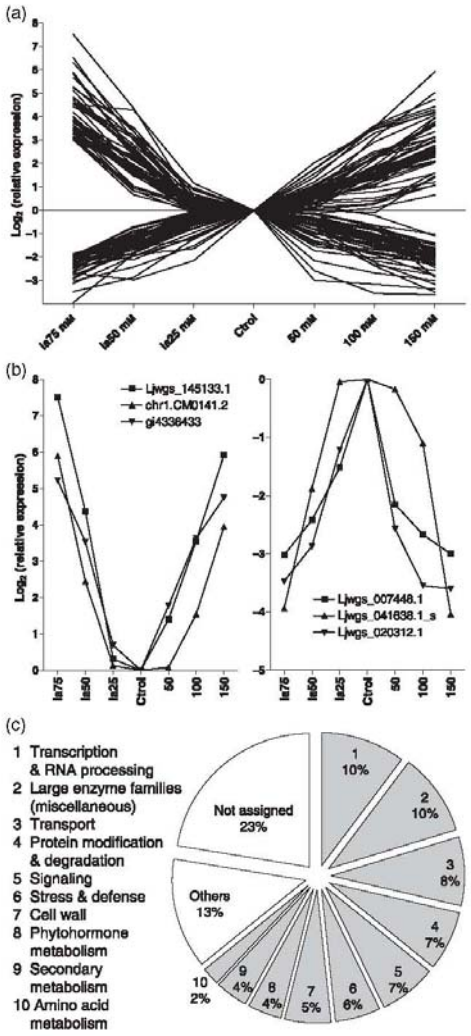


Figure 4. Overview of transcript profiling results.
(a) Expression patterns of the 50 most up-regulated and 50 most down-regulated salt-responsive probesets from microarray profiling experiments. Each expression level (log_2-transformed) represents the mean of three independent experiments. To aid comprehension, error bars were removed and a horizontal line was added at control expression level.
(b) Representative mean expression patterns (log_2-transformed) of six probesets among the 10 most up-regulated and 10 most down-regulated genes.
(c) Non-redundant functional categories of salt-regulated probesets, according to the MapMan software.

index.php. The most abundant transcription factor families were AP2/ERF (24%) and MYB (20%). Both families include characterized members involved in the control of plant

 The Plant Journal, (2008), **53**, 973–987

stress regulons, such as the cold/osmotic stress-induced CBF/DREB sub-family, or the MYB transcription factors implicated in ABA-independent and dependent signalling pathways (Nakano *et al.*, 2006; Yanhui *et al.*, 2006).

Within the AP2/ERF A-5 sub-family of *A. thaliana* transcription factors, no biological function is currently known (Nakano *et al.*, 2006). Therefore, we selected as a target gene the probeset TM0715.25, encoding a putative DREB-like transcription factor similar to those classified in the A-5 sub-family. The transcript levels of this gene were increased by both salt acclimatization and shock, but not by drought, suggesting a role related to ionic but not osmotic stress (Figure 5a). Expression was higher in stems and developing leaves than in mature leaves (Figure 5b), but was induced by salt acclimatization specifically in mature leaves (Figure 5c).

As flavonoid metabolism is involved in salt-stress responses in rice (Walia *et al.*, 2005), we selected another

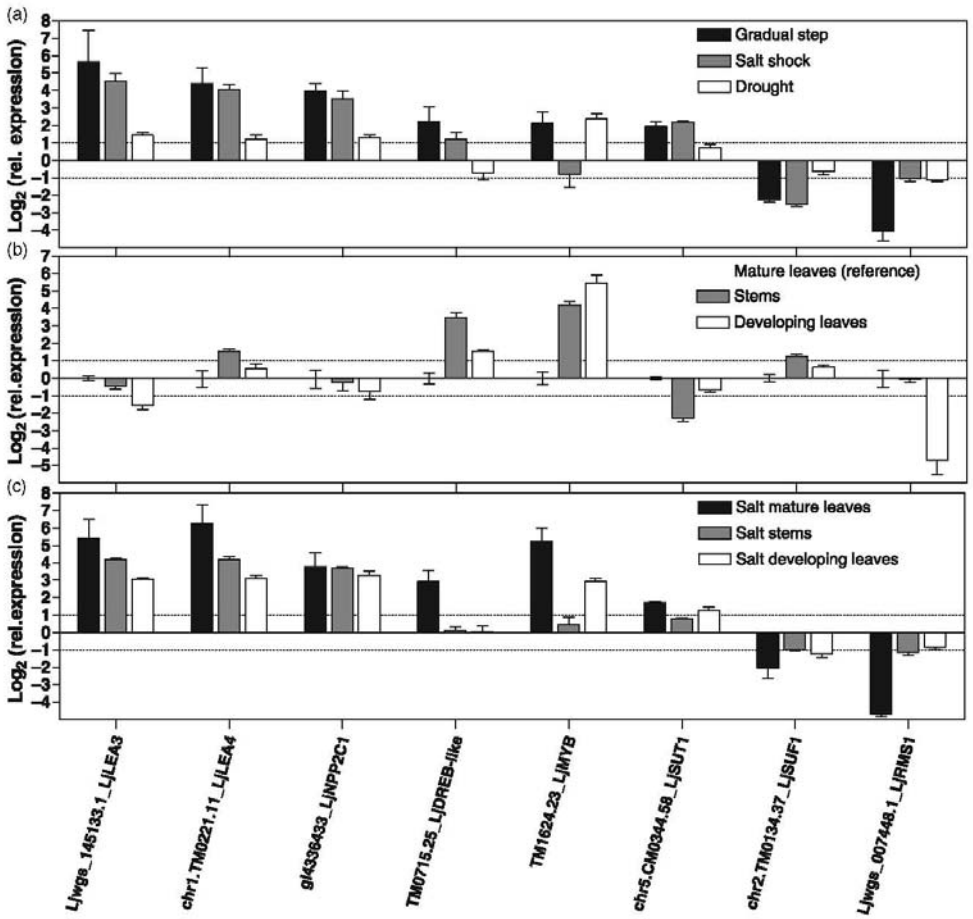


Figure 5. Expression level of selected probesets monitored by quantitative real-time RT-PCR.
Bars represent the mean relative expression level (log_2-transformed) ± SD for three independent replicates. The dotted lines indicate a twofold change. Probesets were sorted according to the respective expression levels in (a).
(a) Gene expression in shoot samples after gradual step acclimatization to 150 mM NaCl (black), *in vitro* salt shock (grey) and drought (white).
(b) Relative gene expression compared to mature leaves, in stems (grey) and developing leaves (white) under control conditions.
(c) Transcriptional regulation after salt acclimatization in mature leaves (black), stems (grey) and developing leaves (white), determined after gradual step acclimatization to 150 mM NaCl and compared to control treatment.

 The Plant Journal, (2008), **53**, 973–987

gene represented by probeset TM1624.23. This putative transcription factor was 79% similar to the *Zea mays* C1 factor protein and 60% similar to AtTT2, both of which are involved in the regulation of pro-anthocyanidin metabolism (Nesi *et al.*, 2001). Expression of this gene was induced by salt acclimatization and drought, but not by salt shock (Figure 5a). It was more highly expressed in stems and developing leaves than in mature leaves, but the salt elicitation was restricted to mature and developing leaves (Figure 5b,c). In addition, several other probesets representing genes putatively involved in the metabolism of phenylpropanoid derivatives were transcriptionally regulated, including homologues of *AtPAL1*, *AtTT4*, *AtTT7* and *AtTT19* (Table S2). We therefore predict a role for TM1624.23 in flavonoid metabolism controlling long-term environmental and developmental responses.

Transport-related genes. Most of the probesets listed in this functional group were putative oligo-peptide, amino acid, sugar or organic acid transporters of the ABC family, suggesting an important role for changes in metabolite allocation under salt acclimatization. As the celery sucrose transporter *AgSUT1* has been reported to be down-regulated under salt stress (Noiraud *et al.*, 2000), we focused on two probesets coding for sucrose transporters: the up-regulated putative *LjSUT1* (probeset chr5.CM0344.58, 80% similar to PsSUT1) and the down-regulated putative *LjSUF1* (probeset chr2.TM0134.37, 85% similar to PvSUF1). The transcript levels of both genes were highly responsive during salt acclimatization and salt shock, but almost non-responsive to drought (Figure 5a). *LjSUT1* was preferentially expressed in leaves compared to stems under control conditions, while the opposite was true for *LjSUF1* (Figure 5b). However, both were elicited under salt stress in all organs, particularly mature leaves (Figure 5c). Because sucrose transporters are involved in loading of sucrose in the phloem (Gottwald *et al.*, 2000), we hypothesize that a complex change in sucrose partitioning between source and sink tissues may be part of the salt-acclimatization process.

No Na^+ transporters or Na^+/H^+ antiporters from the *NHX* or *HKT* families were shown to be up-regulated under salt stress, in contrast to previous results (Maathuis, 2006). The lack of probesets representing orthologous *NHX* genes in the Lotus Genechip® partially explains this discrepancy (only one is present, similar to AtNHX6). However, at least two putative *HKT* genes are represented. The apparent inconsistency may also be explained by our particular choice of long-term acclimatization treatments and restriction to non-lethal stress doses. On the other hand, we found a gene encoding a putative Cl^- channel that was repressed (Ljwgs_016759.2, 86% similar to AtCLC-b, Hechenberger *et al.*, 1996) and identified one putative K^+ transporter as down-regulated by salt treatment (chr5.CM0911.54.1, 79% similar to AtKUP1, Kim *et al.*, 1998). In addition, several genes encoding putative NO_3^- and NH_4^+ transporters exhibited reduced expression, including a NO_3^- transporter that is 71% similar to AtNRT1.2 (probeset chr1.CM0206.164), a NH_4^+ transporter that is 84% similar to AtAMT2 (probesets gi15799271 and Ljwgs_114175.1_s), and a novel NH_4^+ transporter of the LjAMT1 family with 89% similarity to LeAMT1.3 (probesets Ljwgs_028040.1 and Ljwgs_054494.1) (Table S2).

Signalling-related genes. This functional group included putative membrane and cytoplasmatic receptor-like kinases, protein kinases and phosphatases, calmodulin-binding proteins and mitogen-activated protein kinases, suggesting a substantial change of the sensitivity and mode of control exerted by signalling pathways. Homologues of characterized genes integrating environmental signals included the probesets chr4.CM0617.35, encoding a protein 63% similar to the phosphatase type 2C AtABI1 (Leung *et al.*, 1997), and Ljwgs_014485.1 and TM0845.12, which are highly similar to the CBL-interacting protein kinases of *A. thaliana* (Batistic and Kudla, 2004). Among the most up-regulated genes was probeset gi4336433 (Figure 4b), coding for a protein phosphatase type 2C associated with nodule development and highly expressed in nodules and flowers (LjNPP2C1, Kapranov *et al.*, 1999). The expression of *LjNPP2C1* was highly induced by salt acclimatization and salt shock, and to a lesser extent by drought (Figure 5a). Interestingly, transcript levels were equally abundant and elicited in mature leaves, stems and developing leaves (Figure 5b,c). These results may indicate an unexpected link between the nodulation process and the physiology of abiotic stress tolerance. In line with this, other probesets coding for nodulins were transcriptionally regulated, including *LjNOD16*, *LjNOD21*, *LjENOD40-2* and nodulin-like proteins similar to MtN21 and MsENOD8 (Table S2).

Other genes of interest. The stress and defence-related functional group comprised several putative LEA genes, heat shock- and cold- or dehydration-responsive genes. Because LEA and LEA-like genes are among the most up-regulated during stress, we used two highly regulated probesets as positive controls: Ljwgs_145133.1 (LEA group 3) and chr1.TM0221.11 (LEA group 4). Expression of both genes was induced in all tissues tested under salt acclimatization, salt shock and drought (Figure 5a–c).

Among the most down-regulated genes (Figure 4b), the probeset Ljwgs_007448.1 was of particular interest because of the sequence similarity to PsRMS1 and AtMAX4 (89% and 83%, respectively). These orthologues encode the carotenoid-cleaving deoxygenase that is implicated in the control or generation of a long-range transmissible branching signal (Foo *et al.*, 2005). Expression of Ljwgs_007448.1 was low in developing leaves and equally high in stems and mature leaves (Figure 5b). Salt acclimatization resulted in strong reduction, but a similar trend was observed under salt shock

 The Plant Journal, (2008), **53**, 973–987

and drought. Moreover, reduction of expression under salt stress was strongest in mature leaves compared to stems and developing leaves (Figure 5a,c). As *RAMOSUS/MAX* genes interact with auxins and cytokinins for control of morphological patterns, transcriptional down-regulation under stress of this putative *LjRMS1* may indicate a novel link between stress-induced signalling and morphogenetic responses in legumes.

Metabolite profiling

Profiling of soluble metabolites was performed for all treatments in the three independent experiments using gas chromatography/mass spectrometry technology (GC/EI-TOF-MS). Probabilistic principal component analysis (PPCA) was applied to the complete set of observed mass fragments without prior knowledge of metabolite identity (Figure 6a).

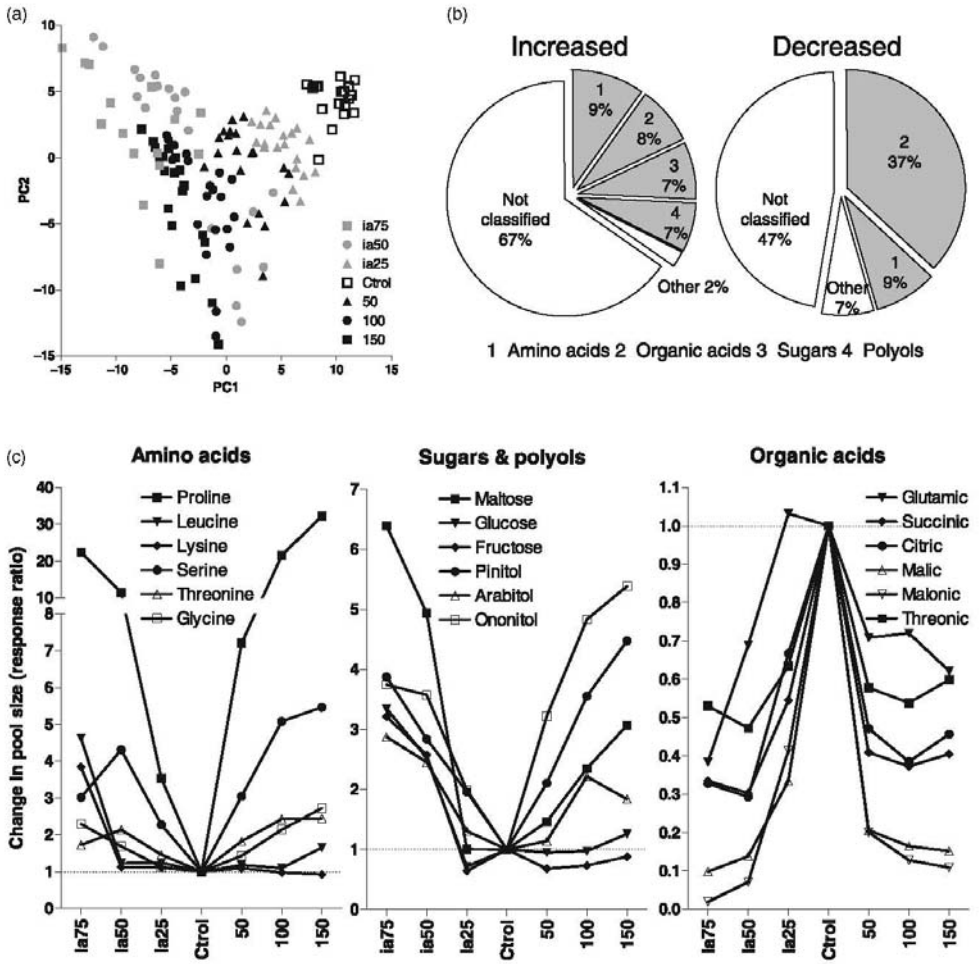


Figure 6. Overview of metabolic profile results.
(a) Metabolic fingerprinting displayed using probabilistic principal component analysis (PPCA). The first two principal components are shown, which comprise the major variation of the data set. Symbols indicate control treatment (open boxes), gradual step acclimatization (black) and initial acclimatization (grey).
(b) Chemical categories for the soluble, salt-responsive metabolites.
(c) Representative patterns of relative pool size changes from the applied chemical categories. Each symbol represents the mean of three independent experiments. Error bars were omitted for improved visualization. The dotted horizontal lines indicate relative control levels.

 The Plant Journal, (2008), **53**, 973–987

This fingerprinting analysis revealed a clear trend of metabolic re-adjustment upon salt acclimatization, illustrated by the first two principal components covering the major variance of the data set. The change in metabolic phenotype was clearly coincident with the increasing salt dose. Sample-to-sample variation, however, also increased substantially at high salt concentrations.

A screen of the metabolic profiles for statistically significant changes (FDR < 0.01) in metabolite pool sizes upon salt acclimatization revealed 147 mass spectral tags; 88 accumulating and 59 decreasing (for definition, see Desbrosses *et al.*, 2005). Due to the restricted availability of reference substances, only approximately one third of the metabolites can currently be identified. As a general metabolic trend, we found a strong decrease in most organic acids and a few amino acids, concomitant with increases of multiple amino acids, sugars, polyols and specific organic acids (Figure 6b). Notable exceptions to these changes were the increase in glucuronic and gulonic acids and the decrease in the primary products of nitrogen assimilation, namely glutamine and asparagine (Table S4). Interestingly, some metabolites related to general stress responses in other species did not accumulate in salt-stressed *L. japonicus* plants. For example, myo-inositol levels did not change, while levels of β-alanine and galactinol were actually slightly decreased (Table S4).

Selected examples of the metabolic changes demonstrated a qualitative similarity to the observed changes in patterns of gene expression (Figure 6c). In most cases, altered metabolites showed strict positive or negative dose dependency, although some changed only at elevated salt concentrations, demonstrating dose-threshold behaviour. Other metabolites appeared to reach an upper or lower pool size, which is particularly evident for most of the organic acids shown in (Figure 6c). This was not caused by analytical limitations resulting from saturation of chromatographic or mass detector capacity.

Discussion

Nutrient homeostasis under salt acclimatization

Salinity-induced nutritional disorders are typically discussed as deficiencies or changes in the requirement for nutrients. However, the influence of salt stress on plant nutrition is highly variable, and depends on the genotype, tissue, growth conditions and chemical characteristics of the soil (Grattan and Grieve, 1999). Exposure of *L. japonicus* plants to increasing concentrations of salts led, as expected, to increases in shoot Na^+ with concomitant decreases in shoot K^+ (Figure 2). Reduction in plant K^+ is the most commonly recognized nutritional change under salt stress, and has been implicated in growth and yield reduction in crops (Grattan and Grieve, 1999). However, many glycophytes are able to substitute Na^+ for K^+ without negative effects on growth (Marschner, 1995). Increases in calcium, magnesium and manganese rule out an NaCl-induced deficiency of these elements in *L. japonicus* shoots under our experimental conditions (Figure 2). The same applies to iron and boron, which appeared to be under homeostatic control. However, we found decreased concentrations of zinc, phosphorus and sulphur under severe salt stress (Figure 2). Salt-induced deficiency of these elements has been reported previously in other species, and could arise from changes in nutrient availability, uptake, transport or partition within the plant (Grattan and Grieve, 1999). As it is generally accepted that an increased supply of nutrients does not necessarily improve the growth of salt-stressed plants under nutrient-sufficient conditions, it is difficult to assess whether a real deficiency exists in our experiments (Grattan and Grieve, 1999). In addition, molybdenum was the most salt-sensitive of all micronutrients in our ionomic profile, decreasing even under mild treatments (Figure 2). To our knowledge, a connection between salinity and Mb content has not been recognized previously. Currently, we have no evidence that salt-acclimatized *L. japonicus* plants reach critical molybdenum deficiency levels, and further analyses will be required to assess this possibility. Interestingly, shoot nitrate reductase activity vital for nitrate assimilation has been shown to decrease under salt stress in tomato (Debouba *et al.*, 2007). As molybdenum is a co-factor of nitrate reductase, our data may provide an explanation for these results.

Salinity-induced changes of gene expression

Major changes in the expression of genes involved in amino acid metabolism as well as nitrogen and organic compound transport, including putative sucrose, amino acid and organic acid transporters, indicated profound changes in central metabolism under long-term salt stress. General metabolic changes were also reflected within the miscellaneous gene group, which included large enzyme families including peroxidases, lipases, glucosidases, glycosyl- and glutathione-*S*-transferases, cytochrome P450s and oxidases, linked to a myriad of cellular processes (Figure S1). Further global control of the acclimatization response is reflected by changes in the transcription and RNA processing, signalling and hormone metabolism categories (Figure 4c).

We identified new molecular candidates that may represent important factors in salt acclimatization in legumes, and further analysed the expression of selected genes within the transcription, transport and signalling functional groups. We also identified a multitude of genes from *L. japonicus* that are homologous to genes that are stress-regulated in other species, not only in the stress and defence-related group but also in other functional categories. For example, many probesets representing genes from secondary metabolism were transcriptionally regulated by salt stress, including

flavonoid, phenylpropanoid and phenol metabolism (Figure S1), which have been previously correlated with biotic and abiotic stress responses (Kliebenstein, 2004; Walia *et al.*, 2005). In addition, we found many transcriptionally regulated genes relating to the cell wall, including expansins, cellulose synthases and glycosyl transferases. In contrast to results reported for rice (Walia *et al.*, 2005), most *L. japonicus* genes in this functional group were down-regulated by salt stress (Table S2 and Figure S1). This observation may be a consequence of the decreased growth and reduced requirement for cell-wall synthesis under long-term salt acclimatization.

Some transcriptional changes were reflected in the metabolomic data. For instance, proline accumulation in plants is common under salt stress (Figure 6c), and seven up-regulated probesets encode proteins that are putatively involved in proline metabolism: Δ-1-pyrroline 5-carboxylase synthetases (P5CS, probesets Ljwgs_006172.2, Ljwgs_032463.1, Ljwgs_053689.1 and chr1.CM0147.99) and Δ-1-pyrroline-5-carboxylate dehydrogenases (P5CDH, probesets chr4.CM0170.37, Ljwgs_019593.1_s and Ljwgs_052588.1_s). We also found induction of two probesets encoding putative myo-inositol-1-phosphate synthases (Ljwgs_091497.1_s and chr4.CM0307.12), concomitant with changes in osmoprotectants from the inositol family (see below). The activity of this enzyme was found to be a rate-limiting step in the biosynthesis of inositol-containing compounds, and is known to be involved in the salt-stress responses of halophytes (Nelson *et al.*, 1998). The depletion of asparagine may be under transcriptional control, as indicated by down-regulation of asparagine synthase 1 (LjAS1, probesets gi897770, gi897770_s and chr5.CM0071.60_s) and up-regulation of two putative asparaginases (Ljwgs_021574.1 and chr5.CM0096.107). In addition, the slight decrease in galactinol was paralleled by down-regulation of a putative galactinol synthase (chr1.CM0122.56).

Approximately one third of the probesets identified in *L. japonicus* showed a significant hit (E value ≤1e-5) when matched to those reported under long-term salt stress in *A. thaliana*, suggesting a certain degree of inter-species similarity in the molecular responses (Table S5) (Sottosanto *et al.*, 2004). Future comparative systems analysis under well-controlled environmental and nutritional conditions will be required to unravel inter-species conservation of salt-stress acclimatization mechanisms.

Salinity-induced changes of the metabolic phenotype

Results from the non-targeted metabolite profiling demonstrated a major and reproducible change of the metabolic phenotype in the course of salt acclimatization, which was most evident for amino acid, sugars and polyols and organic acid metabolism (Figure 6c).

Accumulation of amino acids and other nitrogen-containing compounds is a remarkable biochemical feature of almost all plant stress responses reported so far. This change in nitrogen metabolism has been interpreted as an accumulation of compatible solutes, generation of carbon and nitrogen reserves for future needs, a sink for detoxification of excess nitrogen or for redox potential cycling (Gilbert *et al.*, 1998; Rabie, 1999). This broad response reflects a tightly controlled metabolic shift, and is inconsistent with nitrogen deficiency as a mechanism of salt injury (Grattan and Grieve, 1999). Although it has been extensively shown that nitrogen content may be affected under salinity due to alterations in NO_3^- uptake, non-nodulated salt-stressed legumes exhibited accumulation of NH_4^+ and even increased total nitrogen content (Huq and Larher, 1983; Speer *et al.*, 1994). Therefore, the accumulation of nitrogen-containing compounds may represent a response to a decrease in nitrogen demand caused by reduced growth rates under stress. This contention is in line with the hypothesis of reduced assimilation of inorganic nitrogen in salt-acclimatized *L. japonicus* plants (see below). In addition, the enhanced expression of most of the genes involved in amino acid metabolism (Figure S1) indicates the requirement for *de novo* synthesis already observed in other plant species (Gilbert *et al.*, 1998).

Notable exceptions to the general amino acid behaviour were the two amides glutamine and asparagine. The observed decrease in both glutamine and glutamate may suggest a reduced capacity of NH_4^+ assimilation through the glutamine synthetase/glutamate synthase pathway, as they play a pivotal role in this process (Forde and Lea, 2007). Indeed, these enzymatic activities have been reported to decrease under salinity (Debouba *et al.*, 2007). Although we did not find any regulation of the genes involved in this pathway, we demonstrated consistently reduced expression of two probesets encoding putative nitrite reductases (gi9968472 and chr4.CM0227.40_s, Table S2), which are also involved in nitrogen assimilation, and also down-regulation of genes encoding putative NO_3^- and NH_4^+ transporters. In contrast, the depletion of asparagine was paralleled by transcriptional changes of genes putatively involved in its metabolism (see above). As a major nitrogen transport compound of *L. japonicus* (Waterhouse *et al.*, 1996), the transcriptional control of asparagine levels in the shoot may also reflect decreased inorganic nitrogen assimilation.

Along with nitrogen-containing compounds, sugars and polyols also increase under stress and are known to have protective roles as osmoprotectants (Munns, 2005; Orcutt and Nilsen, 2000). Several compounds related to these chemical groups accumulated in salt-acclimatized *L. japonicus* plants, including the disaccharides maltose and sucrose and the polyols arabitol and erythritol (Figure 6c and Table S4). We also identified the salt-induced methylated inositols pinitol and ononitol, in line with

transcriptional changes of the myo-inositol pathway as described above.

The general depletion in organic acids observed in *L. japonicus* shoots has been demonstrated previously in salt-stressed roots and nodules of the legume alfalfa (Fougere *et al.*, 1991). This phenomenon may reflect an increased energy demand that is met by intensified respiration, carbon allocation to amino acid or sugar pools that are required as compatible solutes, or transport from shoots to roots as carbon source. It is interesting to consider a possible role in compensation for uneven charges entering the plant; organic acids are known to counterbalance unequal uptake of ions as they occur as carboxylic anions under physiological pH (Hinsinger *et al.*, 2003; Marschner, 1995). Moreover, many organic acids accumulate in other species under a variety of stress regimes, such as drought, cold and heat, suggesting that this metabolic change is due to ionic misbalance rather than a decrease in the amount of fixed carbon under stress conditions (Kaplan *et al.*, 2004; Timpa *et al.*, 1986). Notable exceptions to the general depletion of organic acids were glucuronic and gulonic acids (Table S4). As glucuronate is involved in the myo-inositol oxidation pathway that synthesizes nucleotide sugars for cell-wall polysaccharides, the accumulation of glucuronic acid may reflect an inhibition of cell-wall biosynthesis due to decreased growth (Kanter *et al.*, 2005). Alternatively, it might represent an important feature of ascorbic acid metabolism, in line with the parallel increase in gulonic acid. Glucuronic and gulonic acids are intermediates of the uronic pathway that synthesizes ascorbate from myo-inositol (Ishikawa *et al.*, 2006).

Gradual step increase in salts compared to initial acclimatization

The salt-stress physiology of glycophytes is currently interpreted in terms of the biphasic growth model (Munns, 2002, 2005). In essence, this model proposes two growth inhibition phases in response to a gradual increase of salinity. In the first phase, growth inhibition is caused by decreased osmotic potential and thus reduced water availability. After prolonged exposure to salinity, accumulation of ions within plant tissues triggers a second mode of growth inhibition caused by ion toxicity. The biphasic growth model formalizes three essential aspects of salinity: (i) that plant salt-stress physiology ultimately depends on toxicity of ions *per se*, (ii) that the time of exposure is an important variable, and (iii) that appropriate experimentation requires long-term progressive acclimatization to differentiate between (a) hyperosmotic and hyperionic stresses, and (b) physiological acclimatization responses and cellular damage or senescence. In this work, we compared the experimental approach of the biphasic growth model, i.e. the conventional gradual step increase design, to an alternative experimental design based on initial acclimatization. The rationale behind the initial acclimatization is that, given a particular genotype and environment, a range of non-lethal soil salt concentrations should exist that allow the genotype to geminate and develop. Such an experimental design mimics natural and agricultural topsoil-associated salinity, where not only successful plant growth but also establishment on salt-rich soils is required.

Nutrient, transcriptome and metabolome profiling data were used to test whether both experimental designs equally address salt-stress physiology. Firstly, non-supervised analysis suggested that changes do not cluster according to the experimental approach but to the salt-stress dose (Figures 3a and 6a). Secondly, we found only rare statistical evidence for a qualitative differentiation between the acclimatization regimes. For example, under ia50 and ia75 treatments, zinc content decreased while glucose and fructose were increased, a behaviour that is not observed in gradual acclimatization (Figures 2 and 6c). However, it cannot be ruled out that such a difference is due to the differential stress doses perceived by plants between the two designs, as described previously (see Results). Other changes showed only quantitative trends differing between the experimental approaches, with a few changes arising specifically in ia75 plants, probably due to the highly stressful nature of this treatment. Considering the broad scope of the profiling techniques used herein, it could be argued that no major differential effects were observed between the gradual acclimatization and the initial acclimatization experiments. Our results reveal a general equivalence of both salt regimes, supporting the use of the gradual acclimatization approach and the biphasic growth model interpretations, despite previous concerns raised against them (Neumann, 1997).

The stress acclimatization process requires fine tuning of responses

Given that plants acclimatized to a non-lethal salt-stress dose may have reached a stable physiological state essential for survival, it could be argued that most, if not all, of the molecular and metabolic responses reflect the set of plastic physiological changes that, as a whole, allow the plant to cope with the environmental constraint (Lichtenthaler, 1996). If this is the case, how are the various traits regulated and coordinated within the plant in response to changes in the stress dose? From a molecular perspective, a temperature-dependent adjustment has been described for expression of CBF/DREB transcription factors controlling cold-stress responses in *A. thaliana*, demonstrating that stress sensing and response mechanisms are not binary on/off systems (Zarka *et al.*, 2003). Such 'rheostat' control of responses may be interpreted in a simple deterministic model, where all available traits

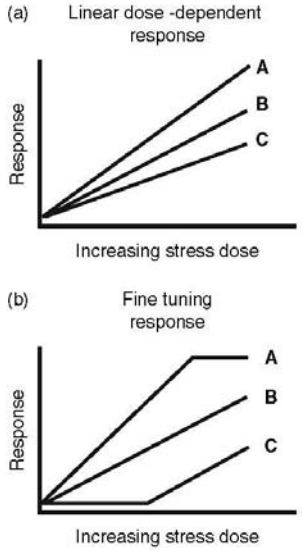


Figure 7. Models of dose-dependent salt-stress acclimatization responses.
(a) Strictly linear dose-dependent responses.
(b) Fine-tuning model, including linear, plateau and threshold dose-dependent responses.
A, B and C represent measurable molecular and metabolic plant traits.

needed to cope with the stress are elicited in a strictly linear dose-dependent manner (Figure 7a). Based on the results of our work, two refinements of this model are necessary (Figure 7b). First, some responses may reach a systems constraint, such as a concentration plateau (trait A). These constraints may be of a genetic or metabolic nature and reflect limitations in the minimal or maximal elicitation of transcript or metabolite pools. Second, some responses are not required at low salt concentrations and will be activated only when a threshold is reached (trait C), suggesting that sensitive responses at the molecular and metabolic systems level are sufficient to compensate for the induced change under low stress intensities. We propose that most major plastic changes under salt acclimatization may be qualitatively explained by this fine tuning model, which includes linear, plateau and threshold dose-dependent responses (Figure 7b). Whether this model may fundamentally be applied to other abiotic stresses remains to be determined.

In conclusion, we performed a systems investigation of salt acclimatization in *L. japonicus*, designed to be as non-biased and comprehensive as possible given the current technological limitations. We found a complex pattern comprising ionomic, transcriptomic and metabolomic responses to salt stress, some of which have not been described in plants before. In addition, we showed that the general transcriptional regulation under long-term salinity is mainly dominated by ion accumulation and toxicity rather than osmotic effects, in line with the proposed physiological biphasic growth model (Munns, 2002, 2005). Finally, we demonstrated the need to refine simple dose-dependence models of salt acclimatization. We now venture to predict that molecular and metabolic differences in acclimatization responses between tolerant and sensitive cultivars may be characterized by three possible features within the framework of the fine-tuning model. Increased tolerance may arise from: (i) changes in the slope of dose-dependent responses, (ii) changes in the upper or lower concentration constraints of a response, or (iii) a shift in the stress dose threshold of factors that do not respond at low stress levels.

Experimental procedures

Plant material, growth conditions, experimental design and sampling

Seeds of *Lotus japonicus* var. Gifu were germinated on agar plates containing agarified half-strength B5 solution (Broughton and Dilworth, 1971) plus 2 mM KNO_3 and 2 mM NH_4NO_3. Four days after imbibition, seedlings were transplanted to soil (Einheit, type null) in 10 cm pots, and irrigated with the above solution. Two salt-stress treatments were implemented: (i) gradual step acclimation to various final NaCl concentrations, or (ii) initial acclimation growth at various NaCl concentrations (see Figure 1a). The gradual acclimation started 8 days post-imbibition, and the salt concentration was increased in three steps of 4 days from 0 to 50, 100 and 150 mM NaCl. A subset of plants was kept at each salt level. Initial acclimation growth exposed plants from germination onwards to nutrient solution supplemented with 25, 50 and 75 mM NaCl. In both cases, fresh nutrient solution was prepared every 4 days. The total duration of greenhouse culture was 28 days under a 16/8 h day/night regime, 23 ± 2°C, 55–65% relative humidity. Whole shoots, excluding cotyledons, were sampled *in situ* into liquid nitrogen in the middle of the light period. Three successive independent experiments (experiments 1, 2 and 3) were performed during the spring season. Each experiment consisted of seven sample sets: control (no salts), 50, 100, 150 (gradual step acclimatization), ia25, ia50 and ia75 (initial acclimatization). Each set had seven independent biological replicate pools of four plants, with the exception of ia75 treatment in experiments 2 and 3 where fewer replicates were available (five and four, respectively). The stress doses used in the experimental designs were not lethal within the cultivation period, and at harvest all plants were in the vegetative stage, and roots did not show nodules. Growth was estimated by determination of mean fresh weight of the pooled shoots, and soil electrical conductivity was determined on 1:2 dried soil:water extraction.

An independent experiment was conducted for verification of the expression patterns of selected target genes, comprising control, gradual acclimatization (150 mM NaCl) and drought-treated plants. In the latter, irrigation was stopped at day 15 after imbibition (the soil water content was 3.79 ± 0.08 and 0.96 ± 0.02 g H_2O g^{-1} dry soil for control and drought treatments, respectively). Whole shoots or separate pools of mature leaves, stems and developing leaves were harvested. In addition, *in vitro* salt-shock experiments were performed using 7-day-old seedlings grown in agarified MS

medium without sucrose (Murashigue and Skoog, 1962). Independent biological replicates, containing at least 50 seedlings each, were sampled 24 h after exposure to 250 mM NaCl together with non-treated controls.

Nutrient profiling analysis

For micro- and macronutrient profiling, 100 mg of plant material was digested with 2 ml HNO_3 (Merck; http://www.merck.de) at 140°C until complete digestion. Then 100 µl of a 100 g l^{-1} LiCl solution (Fluka; http://www.sigmaaldrich.com) was added as a modifier, and the final volume adjusted with ultra-pure water to 10 ml. Element concentrations were determined by inductively coupled plasma/atomic emission spectrometry (ICP-AES) using an IRIS Advantage Duo ER/S (Thermo Fisher; http://www.thermofisher.com). The element emission lines used were: B_2089, Ca_3181, Fe_2599, K_7698, Mg_2790, Mn_2605, Mo_2020, Na_5889, P_1782, S_1820 and Zn_2062. Elemental quantification was validated using IC-CTA-VTL2 Virginia tobacco leaves as a certified reference material (Institute of Nuclear Chemistry and Technology; http://www.ichtj.waw.pl).

Gene expression analysis

Sample tissue for all the biological replicates were pooled to obtain representative RNA for each independent experiment. Total RNA was isolated using the hot borate method (Wan and Wilkins, 1994), and quality and quantity were assessed using a Bioanalyzer-2100 with RNA 6000 NanoChips (Agilent Technologies; http://www.agilent.com) and a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies; http://www.nanodrop.com), respectively. For the microarray experiments, RNA was labelled using the One-Cycle Target labelling kit, hybridized to the Genechip® Lotus1a520343 and scanned, according to the manufacturer's instructions (Affymetrix; http://www.affymetrix.com).

For quantitative real-time RT-PCR analysis, total RNA was DNAse-treated with TURBO DNA-free (Ambion; http://www.ambion.com) and first-strand cDNA was synthesized using an oligo(dT) primer and SuperScript III transcriptase (Invitrogen, http://www.invitrogen.com/). Real-time PCR was performed using 2 × SYBR Green I PCR Mastermix and an ABI Prism 7900HT sequence detection system (Applied Biosystems, http://www.appliedbiosystems.com/). Primer design, reaction conditions, cycling and dissociation-curve parameters, DNA contamination and 3′ to 5′ ratio checks were performed as described by Czechowski *et al.* (2005), without spike-in control and using a reaction volume of 10 µl. Amplification efficiency was assessed using the LinRegPCR program (Ramakers *et al.*, 2003). Analysis of expression data was performed as previously described (Czechowski *et al.*, 2004, 2005), using the geometric mean of four housekeeping genes for normalization (Vandesompele *et al.*, 2002). The housekeeping genes were LjUBQ4 (chr5.CM0956.27), LjGPI-anchored protein (chr3.CM0047.42), LjPP2A (chr2.CM0310.22) and LjUBC10 (chr1.TM0487.4), which were selected from the most stably expressed genes in the plants (Czechowski *et al.*, 2005). A list of all primers used is provided in Table S6.

Metabolic profiling analysis

Frozen plant tissue (60 mg) was extracted using methanol/chloroform, and the polar fraction was prepared by liquid partitioning into water and derivatized (Desbrosses *et al.*, 2005). Gas chromatography coupled to electron impact ionization/time-of-flight mass spectrometry (GC/EI-TOF-MS) was performed using an Agilent 6890N24 gas chromatograph with split or splitless injection connected to a Pegasus III time-of-flight mass spectrometer (LECO Instrumente GmbH; http://www.leco.de) (Wagner *et al.*, 2003). Metabolites were quantified after mass spectral deconvolution (ChromaTOF software version 1.00, Pegasus driver 1.61, LECO) of at least three mass fragments. The peak height representing arbitrary mass spectral ion currents was normalized using the sample fresh weight and ribitol content for internal standardization.

Metabolites were identified using NIST05 software (http://www.nist.gov/srd/mslist.htm) and the mass spectral and retention time index (RI) collection of the Golm metabolome database (Kopka *et al.*, 2005; Schauer *et al.*, 2005). Mass spectral matching was manually checked, and accepted with thresholds of match >650 (maximum 1000) and RI < 1.0 %. RIs represent Kovàts indices (Kovàts, 1958) calculated from additions of C_{12}, C_{15}, C_{19}, C_{22}, C_{32}, C_{36} *n*-alkanes. Table S4 lists not only metabolites identified by standard addition but also mass spectral tags that are as yet unidentified (Desbrosses *et al.*, 2005).

Data analysis and statistics

Hierarchical clustering/supporting trees (HCL-ST) and probabilistic principal component analysis (PPCA) were used as clustering algorithms to analyse data in a non-supervised approach, using TIGR multiple experiment viewer software (TMEV_3.1) and the MetaGeneAlyse webpage (http://metagenealyse.mpimp-golm.mpg.de).

Ionomic and metabolomic data were $\log_{10}$-transformed prior to statistical analysis, which was performed by ANOVA using the following linear model for each of the two experimental designs separately: $y = \beta_0 + \beta_1 t + \beta_2 e + \epsilon$, where y is the measured intensity, t indicates treatment or control, e indicates the experimental block, and ϵ is the error. The P values associated with the null hypothesis H_0: $\beta_1 = 0$ for every nutrient or metabolite were extracted using Student's t-test. A summary P value ($P_{combined}$) for both experimental designs was obtained using H_0: $\beta_{1,step} = 0$ or $\beta_{1,ia} = 0$ and $P_{combined} = 2 \times \min(P_{step}, P_{ia})$. The false discovery rate (FDR) correction (Benjamini and Hochberg, 1995) was applied to the $P_{combined}$ value, and nutrients or metabolites were declared significant if $FDR_{combined} < 0.01$. Both HCL-ST and PPCA were tested for outlier detection, and the biological replicates exp1_ai25_7, exp2_Ctrol_3, exp2_ai50_2, exp2_50_1, exp3_Ctrol_7, exp3_100_7 and exp3_150_7 were omitted from the metabolomic profile (not shown).

Microarray data were analysed using the bioconductor software package for the R programming language (Gentleman *et al.*, 2004). Data quality was assessed using the affy (Gautier *et al.*, 2004) and AffyPLM packages, and expression estimates were obtained using the RMA algorithm (Irizarry *et al.*, 2003). Control and bacterial probesets were removed, and only genes assigned as present ($P < 0.05$) using the MAS5 present/absent algorithm were retained. Statistical testing for differential expression was performed using mixed models with the LIMMA bioconductor package (Smyth, 2004). P values describing control versus treatment comparisons were corrected for multiple testing using the FDR (Benjamini and Hochberg, 1995). Raw data are deposited at Array-Express (http://www.ebi.ac.uk/arrayexpress) as E-MEXP-1204.

Acknowledgements

This work was conducted within the framework of the European LOTASSA project (INCO-CT-2005-517617). We greatly acknowledge the long-standing support of all directors at the Max Planck Institute for Molecular Plant Physiology (MPIMP), the technical assistance of

Ines Fehrle for GC/EI-TOF-MS profiling analyses, and the MPIMP 'greenteam', particularly Britta Hausmann who was in charge of the legume greenhouse. We would like to thank Dr Florian Wagner and the German Resource Center for Genome Research (RZPD, Berlin) team for expert microarray hybridization, Dr Björn Usadel for support with the MapMan software application and for sequence matching, and Dr Ina Talke for support with ICP analysis. D.H.S., F.L., J.K. and M.U. would also like to thank Dr Armin Schlereth for his selfless commitment to our research group and valuable day-to-day assistance.

Supplementary Material

The following supplementary material is available for this article online:

Figure S1. MapMan overview windows for metabolism, large enzyme families (miscellaneous) and transport.

Table S1. Nutrient profile data for all detectable elements.

Table S2. Transcriptomic profile data.

Table S3. Data used to produce (Figure 3c).

Table S4. Metabolic profile data.

Table S5. Sequence matching of *L. japonicus* salt-responsive probesets against *A. thaliana* salt-responsive genes, and list of *L. japonicus* probesets showing a significant hit.

Table S6. Primers used for quantitative real-time RT-PCR.

This material is available as part of the online article from http://www.blackwell-synergy.com

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

References

Achard, P., Cheng, H., De Grauwe, L., Decat, J., Schoutteten, H., Moritz, T., Van Der Straeten, D., Peng, J. and Harberd, N.P. (2006) Integration of plant responses to environmentally activated phytohormonal signals. *Science*, **311**, 91–94.

Batistic, O. and Kudla, J. (2004) Integration and channelling of calcium signaling through the CBL calcium sensor/CIPK protein kinase network. *Planta*, **219**, 915–924.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

Broughton, W.J. and Dilworth, M.J. (1971) Control of leghaemoglobin synthesis in snake beans. *Biochem. J.* **125**, 1075–1080.

Czechowski, T., Bari, R.P., Stitt, M., Scheible, W.R. and Udvardi, M.K. (2004) Real-time RT-PCR profiling of over 1400 Arabidopsis transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *Plant J.* **38**, 366–379.

Czechowski, T., Stitt, M., Altmann, T., Udvardi, M.K. and Scheible, W.R. (2005) Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. *Plant Physiol.* **139**, 5–17.

Debouba, M., Maaroufi-Dghimi, H., Suzuki, A., Ghorbel, H.G. and Gouia, H. (2007) Changes in growth and activity of enzymes involved in nitrate reduction and ammonium assimilation in tomato seedlings in response to NaCl stress. *Ann. Bot.* **99**, 1143–1151.

Desbrosses, G.G., Kopka, J. and Udvardi, M.K. (2005) *Lotus japonicus* metabolic profiling. Development of gas chromatography–mass spectrometry resources for the study of plant–microbe interactions. *Plant Physiol.* **137**, 1302–1318.

Foo, E., Bullier, E., Goussot, M., Foucher, F., Rameau, C. and Beveridge, C.A. (2005) The branching gene RAMOSUS1 mediates interactions among two novel signals and auxin in pea. *Plant Cell*, **17**, 464–474.

Forde, B.G. and Lea, P.J. (2007) Glutamate in plants: metabolism, regulation, and signalling. *J. Exp. Bot.* **58**, 2339–2358.

Fougere, F., Le Rudulier, D. and Streeter, J.G. (1991) Effects of salt stress on amino acid, organic acid and carbohydrate composition of roots, bacteroids and cytosol in alfalfa (*Medicago sativa* L.). *Plant Physiol.* **96**, 1228–1236.

Gautier, L., Cope, L., Bolstad, B.M. and Irizarry, R.A. (2004) Affy-analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.

Gentleman, R.C., Carey, V.J., Bates, D.M. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80.

Gilbert, G.A., Gadush, M.V., Wilson, C. and Madore, M.A. (1998) Amino acid accumulation in sink and source tissues of *Coleus blumei* Benth. during salinity stress. *J. Exp. Bot.* **49**, 107–114.

Gottwald, J.R., Krysan, P.J., Young, J.C., Evert, R.F. and Sussman, M.R. (2000) Genetic evidence for the in planta role of phloem-specific plasma membrane sucrose transporters. *Proc. Natl Acad. Sci. USA*, **97**, 13979–13984.

Graham, P.H. and Vance, C.P. (2003) Legumes: importance and constraints to greater use. *Plant Physiol.* **131**, 872–877.

Grattan, S.R. and Grieve, C.M. (1999) Mineral nutrient acquisition and response by plants grown in saline environments. In *Handbook of Plant and Crop Stress* (Pessarakli, M., ed.). New York: Marcel Dekker Inc., pp. 203–229.

Hechenberger, M., Schwappach, B., Fischer, W.N., Frommer, W.B., Jentsch, T. and Steinmeyer, K. (1996) A family of putative chloride channels from *Arabidopsis* and functional complementation of a yeast strain with a CLC gene disruption. *J. Biol. Chem.* **271**, 33632–33638.

Hinsinger, P., Plassard, C., Tang, C. and Jaillard, B. (2003) Origins of root-mediated pH changes in the rhizosphere and their responses to environmental constraints: a review. *Plant Soil*, **248**, 43–59.

Hose, E., Clarkson, D.T., Steudle, E., Schreiber, L. and Hartung, W. (2001) The exodermis: a variable apoplastic barrier. *J. Exp. Bot.* **52**, 2245–2264.

Huq, S.M.I. and Larher, F. (1983) Osmoregulation in higher plants: effects of NaCl salinity on non-nodulated *Phaseolus aureus* L. *New Phytol.* **93**, 209–216.

Irizarry, R., Bolstad, B., Collin, F., Cope, L., Hobbs, B. and Speed, T. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15.

Ishikawa, T., Dowdle, J. and Smirnoff, N. (2006) Progress in manipulating ascorbic acid biosynthesis and accumulation in plants. *Physiol. Plant.* **126**, 343–355.

Kanter, U., Usadel, B., Guerineau, F., Li, Y., Pauly, M. and Tenhaken, R. (2005) The inositol oxygenase gene family of *Arabidopsis* is involved in the biosynthesis of nucleotide sugar precursors for cell-wall matrix polysaccharides. *Planta*, **221**, 243–254.

Kaplan, F., Kopka, J., Haskell, D.W., Zhao, W., Schiller, K.C., Gatzke, N., Sung, D.Y. and Guy, C.L. (2004) Exploring the temperature-stress metabolome of Arabidopsis. *Plant Physiol.* **136**, 4159–4168.

Kapranov, P., Jensen, T.J., Poulsen, C., De Bruijin, F.J. and Szyglowski, K. (1999) A protein phosphatase 2C gene, LjNPP2C1, from *Lotus japonicus* induced during root nodule development. *Proc. Natl Acad. Sci. USA*, **96**, 1738–1743.

Kim, E.J., Kwak, J.M., Uozumi, N. and Schroeder, J.I. (1998) AtKUP1: an *Arabidopsis* gene encoding high-affinity potassium transport activity. *Plant Cell*, **10**, 51–62.

Kliebenstein, D.J. (2004) Secondary metabolites and plant/environment interactions: a view through *Arabidopsis* tinged glasses. *Plant Cell Environ.* **27**, 675–684.

Kopka, J., Schauer, N., Krueger, S. *et al.* (2005) GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics*, **21**, 1635–1638.

Kovàts, E.S. (1958) Gas-chromatographische Charakterisierung organischer Verbindungen: Teil 1. Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. *Helv. Chim. Acta*, **41**, 1915–1932.

Leung, J., Merlot, S. and Giraudat, J. (1997) The *Arabidopsis* ABSCISIC-ACID-INSENSITIVE2 (ABI2) and ABI1 genes encode homologous protein phosphatases 2C involved in abscisic acid signal transduction. *Plant Cell*, **9**, 759–771.

Lichtenthaler, H.K. (1996) Vegetation stress: an introduction to the stress concept in plants. *J. Plant Physiol.* **148**, 4–14.

Marschner, H. (1995) *Mineral Nutrition of Higher Plants*. London: Academic Press.

Maathuis, F.J.M. (2006) The role of monovalent cation transporters in plant responses to salinity. *J. Exp. Bot.* **57**, 1137–1147.

Munns, R. (2002) Comparative physiology of salt and water stress. *Plant Cell Environ.* **25**, 239–250.

Munns, R. (2005) Genes and salt tolerance: bringing them together. *New Phytol.* **167**, 645–663.

Murashigue, T. and Skoog, F. (1962) A revised medium for rapid growth and bioassays with tobacco tissue cultures. *Physiol. Plant.* **15**, 473–497.

Nakano, T., Suzuki, K., Fujimura, T. and Shinshi, H. (2006) Genome-wide analysis of ERF gene family in Arabidopsis and rice. *Plant Physiol.* **140**, 411–432.

Nelson, D.E., Rammesmayer, G. and Bohnert, H.J. (1998) Regulation of cell-specific inositol metabolism and transport in plant salinity tolerance. *Plant Cell*, **10**, 753–764.

Nesi, N., Jond, C., Debeaujon, I., Caboche, M. and Lepiniec, L. (2001) The Arabidopsis TT2 gene encodes an R2R3 MYB domain protein that acts as a key determinant for proanthocyanidin accumulation in developing seed. *Plant Cell*, **13**, 2099–2114.

Neumann, P. (1997) Salinity resistance and plant growth revisited. *Plant Cell Environ.* **20**, 1193–1198.

Noiraud, N., Delrot, S. and Lemoine, R. (2000) The sucrose transporter of celery. Identification and expression during salt stress. *Plant Physiol.* **122**, 1447–1455.

Orcutt, D.M. and Nilsen, E.T. (2000) *The Physiology of Plants under Stress: Soil And Biotic Factors*. New York: John Wiley & Sons Inc.

Rabie, E. (1999) Altered nitrogen metabolism under environmental stress conditions. In *Handbook of Plant and Crop Stress* (Pessarakli, M., ed.). New York: Marcel Dekker Inc., pp. 349–363.

Ramakers, C., Ruijter, J.M., Lekanne Deprez, R.H. and Moorman, A.F.M. (2003) Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. *Neurosci. Lett.* **339**, 62–66.

Schauer, N., Steinhauser, D., Strelkov, S. *et al.* (2005) GC–MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett.* **579**, 1332–1337.

Smyth, G.K. (2004) Linear models and empirical Bayes for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, No. 1, article 3.

Sottosanto, J.B., Gelli, A. and Blumwald, E. (2004) DNA array analyses of *Arabidopsis thaliana* lacking a vacuolar Na^+/H^+ antiporter: impact of AtNHX1 on gene expression. *Plant J.* **40**, 752–771.

Speer, M., Brune, A. and Kaiser, W.M. (1994) Replacement of nitrate by ammonium as the nitrogen source increases the salt sensitivity of pea plants. I. Ion concentrations in roots and leaves. *Plant Cell Environ.* **17**, 1215–1221.

Teakle, N.L., Real, D. and Colmer, T.D. (2006) Growth and ion relations in response to combined salinity and waterlogging in the perennial forage legumes *Lotus corniculatus* and *Lotus tenuis*. *Plant Soil*, **289**, 369–383.

Tester, M. and Davenport, R. (2003) Na^+ tolerance and Na^+ transport in plants. *Ann. Bot.* **91**, 503–527.

Timpa, J.D., Burke, J.J., Quisenberry, J.E. and Wendt, C.W. (1986) Effects of water stress on the organic acid and carbohydrate compositions of cotton plants. *Plant Physiol.* **82**, 724–728.

Usadel, B., Nagel, A., Thimm, O. *et al.* (2005) Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of corresponding genes, and comparison with known responses. *Plant Physiol.* **138**, 1195–1204.

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A. and Speleman, F. (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* **3**, research0034.1–0034.11.

Wagner, C., Sefkow, M. and Kopka, J. (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry*, **62**, 887–900.

Walia, H., Wilson, C., Condamine, P. *et al.* (2005) Comparative transcriptional profiling of two contrasting rice genotypes under salinity stress during the vegetative growth stage. *Plant Physiol.* **139**, 822–835.

Wan, C.Y. and Wilkins, T.A. (1994) A modified hot-borate method significantly enhances the yield of high-quality RNA from cotton *Gossypium hirsutum* L. *Anal. Biochem.* **223**, 7–12.

Waterhouse, R.N., Smyth, A.J., Massonneau, A., Prosser, I.M. and Clarkson, D.T. (1996) Molecular cloning and characterization of asparagine synthetase from *Lotus japonicus*: dynamics of asparagine synthesis in N-sufficient conditions. *Plant Mol. Biol.* **30**, 883–897.

Yanhui, C., Xiaoyuan, Y., Kun, H. *et al.* (2006) The MYB transcription factor superfamily of Arabidopsis: expression analysis and phylogenetic comparison with the rice MYB family. *Plant Mol. Biol.* **60**, 107–124.

Zarka, D.G., Vogel, J.T., Cook, D. and Thomashow, M.F. (2003) Cold induction of Arabidopsis CBF genes involved multiple ICE (inducer of CBF expression) promoter elements and a cold-regulatory circuit that is desensitized by low temperature. *Plant Physiol.* **133**, 910–918.

 The Plant Journal, (2008), **53**, 973–987

[Plant Signaling & Behavior 3:8, 1-3; August 2008]; ©2008 Landes Bioscience

Article Addendum

Metabolome-ionome-biomass interactions

What can we learn about salt stress by multiparallel phenotyping?

Diego H. Sanchez,[1,†] Henning Redestig,[1,†] Ute Krämer,[2] Michael K. Udvardi[3] and Joachim Kopka[1,*]

[1]Max Planck Institute for Molecular Plant Physiology; Wissenschaftspark Golm; Potsdam-Golm, Germany; [2]Bioquant Center; University of Heidelberg; Heidelberg, Germany; [3]Samuel Roberts Noble Foundation; Ardmore, Oklahoma USA

[†]These authors contributed equally to this work.

Abbreviations: ANOVA, analysis of variance; GC/EI-TOF-MS, gas chromatography—electron impact ionization—time of flight—mass spectrometry; ICP-AES, inductively coupled plasma—atomic emission spectrometry

Key words: acclimation, ionomic, *lotus*, metabolic, metabolomic, nutrients, salinity, salt stress

Long-term exposure of plants to saline soil results in mineral ion imbalance, altered metabolism and reduced growth. Currently, the interaction between ion content and plant metabolism under salt-stress is poorly understood. Here we present a multivariate correlation study on the metabolome, ionome and biomass changes of *Lotus japonicus* challenged by salt stress. Using latent variable models, we show that increasing salinity leads to reproducible changes of metabolite, ion and nutrient pools. Strong correlations between the metabolome and the ionome or biomass may allow one to estimate the degree of salt stress experienced by a plant based on metabolite profiles. Despite the apparently high predictive power of the models, it remains to be investigated whether such metabolite profiles of non- or moderately-stressed plants can be used by breeding programs as ideal ideotypes for the selection of enhanced salt-tolerant genotypes.

Acclimation of plants to saline soils involves changes in the uptake, transport and/or partitioning of mineral ions.[1-3] These responses not only alter ion concentrations but also impair metabolism and growth.[4] Exactly how metabolism as a whole changes in response to salinity is still unknown because of the complexity of the processes involved. Nevertheless, one might expect plant metabolism to respond in a predictable way to salt stress. With this in mind we carried out a multivariate correlation analysis of 137 metabolome and ionome profiles, and the corresponding biomass measurements, of shoot samples from *Lotus japonicus* exposed to two different salinity regimes.[5]

*Correspondence to: Joachim Kopka; Max Planck Institute for Molecular Plant Physiology; Wissenschaftspark Golm; Am Mühlenberg 1; Potsdam-Golm 14476 Germany; Tel.: +49.331.567.8210; Email: kopka@mpimp-golm.mpg.de

Submitted: 05/28/08; Accepted: 05/28/08

Previously published online as a *Plant Signaling & Behavior* E-publication: http://www.landesbioscience.com/journals/psb/article/6347

Addendum to: Sanchez DH, Lippold F, Redestig H, Hannah M, Erban A, Krämer U, Kopka J, Udvardi MK. Integrative functional genomics of salt acclimatization in the model legume *Lotus japonicus*. Plant J 2008; doi:10.1111/j.1365-313X.2007.03381.x.

Metabolome data obtained using GC/EI-TOF-MS technology were analyzed using the TagFinder software,[6] resulting in a series of discrete metabolic-features. A metabolic-feature may be defined to represent a quantitative signal, measured by any analytical means or technology, which is distinct from analytical signals that arise as artefacts from electronic or chemical noise. A total of 1019 metabolic-features were obtained after filtering for those represented by 3 or more inter-correlated mass fragments.[6] Corresponding ionomic data were obtained using ICP-AES technology and included measurements of Na and 10 macro- and micro-nutrients (K, Ca, S, P, Mg, B, Mn, Fe, Zn and Mo).[5]

To integrate metabolomic and ionomic data, we used the statistical multivariate regression technique called orthogonal projections to latent structures (OPLS[7,8]), which performs a regression of two matrices or a matrix versus a single variable and simultaneously corrects the resulting model for systematic, irrelevant variance. The metabolite profile matrix was regressed in three different models against the concentrations of Na, K and a matrix of all nutrients excluding K and Na. These regressions were designated metabolome-[Na], metabolome-[K] and metabolome [nutrients-K] models, respectively. With OPLS it is possible to estimate how well associated the different metabolites are with the modeled variance of the different ions. The used measure is called the correlation loading and can be interpreted as a multivariate version of the standard Pearson correlation. In order to compare how the metabolic profiles may predict the different matrices, we regressed the correlation loadings vectors of the three different models amongst each other (Fig. 1). Remarkably, the loadings were highly correlated. Despite the high magnitude of change in K content compared to the other nutrients under salinity, the metabolome-[K] and metabolome-[nutrients-K] loadings were nearly identical, highlighting that K levels correlate strongly with the main metabolome correlated variance in the rest of the nutrient matrix. These observations suggest that salinity leads to reproducible changes in metabolite pools which match both the concentration of salt accumulated in the shoot and induced changes in the content of other elements.[5] Since metabolome profiles have been considered a predictor of plant biomass under non-stressed growth conditions,[9] a metabolome-biomass model was evaluated

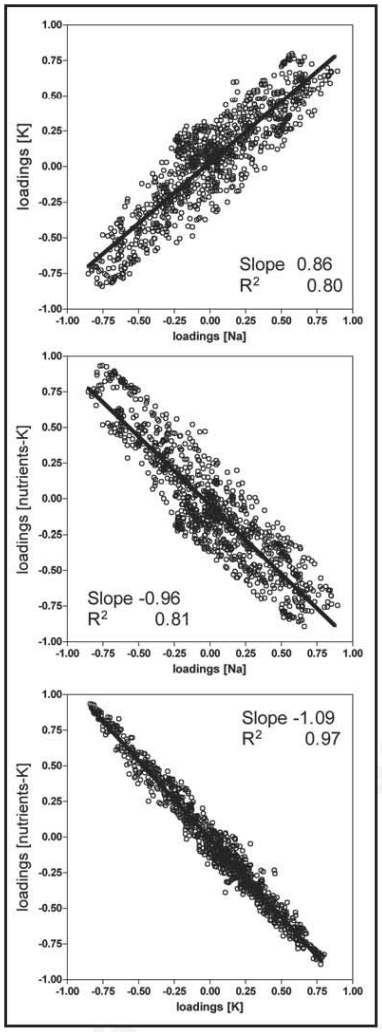


Figure 1. Regression of the correlation loadings obtained from the models metabolome [Na], metabolome [K] and metabolome [nutrients-K].

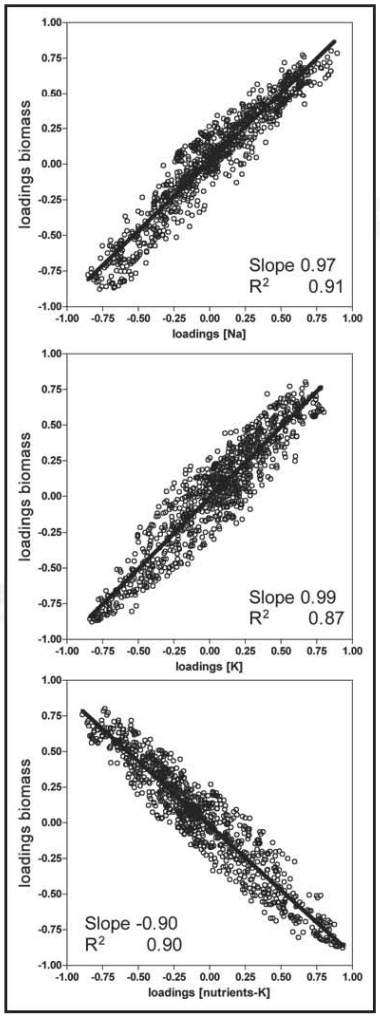


Figure 2. Regression of the correlation loadings of the metabolome [Na], metabolome [K] and metabolome [nutrients-K] models with the metabolome biomass model.

for the stress cue of our experimental setup. The metabolome data appeared to be correlated to shoot biomass in a manner similar to the predictability of [Na], [K] and [nutrients-K] (Fig. 2). Presumably, this observation reflects the property of the plant system to integrate in a highly interdependent process the nutritive elements, metabolism and growth.

The correlation loadings of the models allowed a ranking of metabolite-features according to their contribution to the modeled regressions. We used the magnitude of the weight of each metabolic-feature to assess which metabolites may be more characteristic or diagnostic of salt stress, as determined by Na levels (Table 1). Note that among the positive and negative most predictive metabolic-features, only 4 could be chemically identified so far: gulonic acid, glucuronic acid, β-alanine and cis-cinnamic acid. In addition, models based on the metabolic profiles appeared to be robust predictors of salt content or biomass (Fig. 3).

Although correlation per se does not reveal causality, our analysis suggests that salt stress-induced changes in shoot metabolites represent an integrative systems response which links salt accumulation and altered ion balance to the control of growth and final biomass.

Table 1 The top-most positively and negatively correlated metabolites of the metabolite [Na] regression model

Feature	Metabolite	loading [Na]
3498	Gulonic acid	0.849290866
3500	UNKNOWN	0.828822131
3518	UNKNOWN	0.811654749
5134	UNKNOWN	0.801678882
4338	A177004	0.784970498
2791	A140003	0.776360031
2354	Glucuronic acid	0.772587791
7019	A197007	0.759543257
5211	A143004	0.732318855
538	A211001	0.723284198
1528	A144003	-0.75534488
1526	A144003	-0.77016393
5438	Alanine, beta-	-0.7776222
4752	A158003	-0.79205792
3551	A161003	-0.79236556
3027	UNKNOWN	-0.79750797
3012	A154002	-0.7999629
3021	UNKNOWN	-0.8094395
3016	UNKNOWN	-0.81235463
2776	Cinnamic acid, 4-hydroxy-, cis-	-0.82861256

Un-identified metabolites that have been detected before are denoted by a Golm Metabolite Database code,[10] while UNKNOWN metabolic-features are yet to be archived in the database.

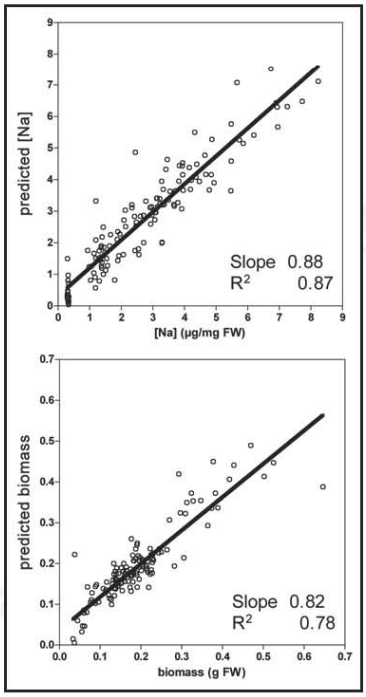


Figure 3. Predictive power of the analysis, as revealed by a linear regression between the measured [Na] or biomass and the predicted [Na] or biomass from the models. The predictions were performed based on 10-fold cross-validation, where in each segment the true values of Na content and biomass were held out and predicted from the corresponding metabolome data using the OPLS model.

Since accumulation of salts and ion toxicity within the plant must be considered the primary cause of growth inhibition and senescence under long-term salt stress,[11] the high predictive qualities of models based on metabolome phenotyping may allow the estimation of the degree of salt stress experienced by a plant. Thus, it may be possible in future to use metabolic fingerprinting as a breeding tool to select individual plants that best cope with salt stress. On the other hand, given the interdependent nature of plant responses to environmental stress, metabolite-based models may not reveal unique properties of salt accumulation or reduced growth. Due to the high diversification of biosynthetic capabilities, the transfer of knowledge between species belonging to different plant clades may be restricted to the conserved metabolic responses.[4]

Acknowledgements

This work was conducted within the framework of the European LOTASSA project (INCO-CT-2005-517617).

References

1. Marschner H. Mineral nutrition of higher plants, 2nd ed. London: Academic Press Limited 1995.
2. Grattan SR, Grieve CM. Mineral nutrient acquisition and response by plants grown in saline environments. In: Pessarakli M, ed. Handbook of plant and crop stress 2nd ed. New York: Marcel Dekker Inc 1999; 203-29.
3. Tester M, Davenport R. Na^+ tolerance and Na^+ transport in plants. Ann Bot 2003; 91:503-27.
4. Sanchez DH, Siahpoosh MR, Roessner U, Udvardi MK, Kopka J. Plant metabolomics reveals conserved and divergent metabolic responses to salinity. Physiol Plant 2008; 132:209-19.
5. Sanchez DH, Lippold F, Redestig H, Hannah M, Erban A, Krämer U, Kopka J, Udvardi MK. Integrative functional genomics of salt acclimatization in the model legume *Lotus japonicus*. Plant J 2008; 53:973-87.
6. Luedemann A, Strassburg K, Erban A, Kopka J. TagFinder for the quantitative analysis of gas chromatography-mass spectrometry (GC-MS) based metabolite profiling experiments. Bioinformatics 2008; 24:732-7.
7. Trygg J, Wold S. Orthogonal projections to latent structures (O-PLS). J Chemom 2002; 16:119-28.
8. Bylesjö M, Eriksson D, Kusano M, Moritz T, Trygg J. Data integration in plant biology: the O2PLS method for combined modelling of transcript and metabolite data. Plant J 2007; 52:1181-91.
9. Meyer RC, Steinfath M, Lisec J, Becher M, Witucka-Wall H, Torjek O, Fiehn O, Eckardt A, Willmitzer L, Selbig J, Altmann T. The metabolic signature related to high plant growth rate in *Arabidopsis thaliana*. Proc Nat Acad Sci USA 2007; 104:4759-64.
10. Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmueller E, Doermann P, Weckwerth W, Gibon Y, Stitt M, Willmitzer L, Fernie AR, Steinhauser D. GMD@CSB.DB: the Golm Metabolome Database. Bioinformatics 2005; 21:1635-8.
11. Munns R. Comparative physiology of salt and water stress. Plant Cell Environ 2002; 25:239-50.